Anales de la Ciencia

LA VERDAD PASA

¿Hay algo equivocado con respecto al método científico?

Por Jonah Lehrer

13 de diciembre de 2010

El 18 de setiembre de 2007 unas pocas docenas de neurocientíficos, psiquiatras y ejecutivos de compañías farmacéuticas se reunieron en el salón de conferencias de un hotel en Bruselas para escuchar algunas noticias sorprendentes. Tenían que ver con una clase de fármacos conocidos como atípicos o anti sicóticos de segunda generación que aparecieron en el mercado a principios de los noventas. Los fármacos, vendidos bajo nombres de marcas tales como Abilify, Seroquel y Zyprexa, habían sido probados en esquizofrénicos en varias pruebas de clínicas grandes, todas las cuales habían demostrado una disminución dramática en los síntomas psiquiátricos de los sujetos. Como resultado, los anti sicóticos de segunda generación se habían convertido en una de las clases farmacéuticas de más rápido crecimiento y de las más rentables. Para el 2001, el Zyprexa de Eli Lilly estaba generando más ganancias que el Prozac. Sigue siendo el fármaco de mayor venta de la compañía.

Pero la información presentada en la reunión en Bruselas puso en claro que algo extraño estaba sucediendo: el poder terapéutico de los fármacos parecía estar disminuyendo continuamente. Un estudio reciente mostró un efecto que era menos de la mitad de aquel efecto documentado en las primeras pruebas a comienzos de los años noventas. Muchos investigadores comenzaron a argumentar que aquellos fármacos tan caros no eran en nada mejores que los anti sicóticos de primera generación, los cuales habían estado en uso desde los años cincuenta. “De hecho, a veces ahora se ven incluso peores,” me dijo John Davis, un profesor de psiquiatría en la Universidad de Illinois en Chicago.

Antes que pueda confirmarse la efectividad de un fármaco debe ser probado una y otra vez. Diferentes científicos en laboratorios diferentes necesitan repetir los protocolos y publicar sus resultados. La prueba de la réplica, como se le conoce, es el fundamento de la investigación moderna. La réplica es como la comunidad se hace respetar a sí misma. Es una salvaguarda contra la adulación de la subjetividad. La mayor parte del tiempo los científicos saben qué resultados quieren, y eso puede influenciar los resultados que obtienen. La premisa de la replicabilidad es para que la comunidad científica pueda corregir estos defectos.

Pero ahora toda clase de descubrimientos bien establecidos y confirmados de muchas maneras han comenzado a verse con una creciente incertidumbre. Es como si nuestros hechos estuviesen perdiendo su veracidad: afirmaciones que han sido consagradas en los libros de texto de pronto se han vuelto no demostrables. Este fenómeno todavía no tiene un nombre oficial, pero está sucediendo en un amplio rango de campos, desde la psicología hasta la ecología. En el campo de la medicina el fenómeno parece extremadamente extendido, afectando no solamente a los anti sicóticos sino también las terapias desde los ***stents*** cardíacos, pasando por la Vitamina E y los antidepresivos: Davis tiene un análisis que está por aparecer demostrando que la eficacia de los antidepresivos ha disminuido tanto como hasta el triple en décadas recientes.

Para muchos científicos, el efecto es especialmente perturbador debido a lo que expone con respecto al proceso científico. Si la replicación es lo que separa al rigor de la ciencia de la fangosidad de la pseudociencia, ¿dónde colocamos todos estos descubrimientos rigurosamente validados que ya no pueden probarse? ¿Cuáles resultados deben creerse? Francis Bacon, el precoz filósofo moderno y pionero del método científico, declaró una vez que los experimentos eran esenciales, porque nos permitían "plantearle preguntas a la naturaleza." Pero parece que la naturaleza a menudo nos da respuestas diferentes.

Jonathan Schooler era un joven estudiante graduado en la Universidad de Washington en los años ochenta cuando descubrió un nuevo hecho sorprendente sobre el lenguaje y la memoria. En ese tiempo, se creía de manera generalizada que el acto de describir nuestros recuerdos los mejoraba. Pero, en una serie de ingeniosos experimentos, Schooler demostró que a los sujetos a quienes se les mostraba un rostro y luego se les pedía que lo describieran tenían menos probabilidades de reconocer el rostro cuando se les mostraba más tarde que aquellos que simplemente lo habían visto. Schooler le llamó a este fenómeno "oscurecimiento verbal."

El estudio le convirtió en una estrella académica. Desde su publicación inicial en 1990, ha sido citado más de cuatrocientas veces. Antes de mucho, Schooler había extendido el modelo a una variedad de tareas diferentes, tales como recordar el sabor de un vino, identificar la mejor jalea de fresa y resolver difíciles rompecabezas creativos. En cada ejemplo, pedirle a la gente que pusiera sus percepciones en palabras condujo a dramáticas disminuciones en el desempeño.

Pero mientras Schooler estaba dando a conocer estos resultados en publicaciones de elevada reputación, una preocupación secreta le roía por dentro: se estaba mostrando que era difícil replicar sus primeros descubrimientos. "A menudo todavía veía un efecto, pero el efecto simplemente no era tan fuerte," me dijo. "Era como si el oscurecimiento verbal, mi gran nueva idea, se estuviese haciendo más débil." Al principio, asumió que había cometido un error en el diseño experimental o un cálculo estadístico equivocado. Pero no podía encontrar nada mal en su investigación. Entonces concluyó en que su lote inicial de sujetos bajo investigación debió haber sido excepcionalmente susceptible al

oscurecimiento verbal. (De igual manera, John Davis ha especulado que parte de la caída en la efectividad de los anti sicóticos puede ser atribuida al uso de sujetos que sufren de formas más ligeras de psicosis lo que los hace menos probables de mostrar un mejoramiento dramático.) “No fue una explicación muy satisfactoria,” dice Schooler. “Uno de mis mentores me dijo que mi verdadero error era el de tratar de replicar mi trabajo. Me dijo que al hacer eso sólo me estaba predisponiendo a la decepción.”

Schooler trató de sacarse el problema de su mente; sus colegas le aseguraron que esas cosas pasaban todo el tiempo. En los años siguientes, encontró nuevos motivos de investigación, se casó y tuvo hijos. Pero su problema de replicación se fue haciendo cada vez peor. Su primer intento de replicar el estudio de 1990, en 1995, resultó en un efecto que era treinta por ciento más pequeño. Al año siguiente, la magnitud del efecto se redujo otro treinta por ciento. Cuando otros laboratorios repitieron los experimentos de Schooler, obtuvieron un margen similar de datos, con una tendencia distintiva a la baja. “Esto fue profundamente frustrante,” dice. “Era como si la naturaleza me hubiese dado este gran resultado y luego tratara de quitármelo.” En privado, Schooler comenzaba a referirse al problema como “habituación cósmica,” por analogía a la disminución en la respuesta que ocurre cuando los individuos se habitúan a un estímulo particular. “La habituación es la razón por la cual usted no nota aquella cosa que siempre está allí,” dice Schooler. “Es un proceso inevitable de ajuste, un apalancamiento del ánimo. Comencé a bromear que era como si el cosmos se estuviera habituando a mis ideas. Lo tomé muy personalmente.”

Schooler es ahora un profesor con titularidad en la Universidad de California en Santa Bárbara. Tiene el cabello negro y crespo, ojos de color verde pálido y el porte relajado de alguien que vive a cinco minutos de su playa favorita. Cuando habla tiende a distraerse por sus propias digresiones. Podría comenzar con un punto acerca de la memoria, lo que le recuerda una cita favorita de William James, lo cual inspira un extenso soliloquio sobre la importancia de la introspección. Antes de mucho, estaremos viendo imágenes de ***El Hombre en Llamas*** en su iPhone, lo cual nos lleva de vuelta a la frágil naturaleza de la memoria.

Aunque el oscurecimiento verbal sigue siendo una teoría ampliamente aceptada – a menudo se le invoca en el contexto del testimonio ocular, por ejemplo – Schooler todavía se encuentra un poco fastidiado en el cosmos. “Sé que ya debí haber avanzado,” dice. “Realmente debería dejar de hablar de esto. Pero no puedo.” Eso es porque está convencido de que ha tropezado y caído en un problema serio, uno que aflige a muchas de las más emocionantes ideas nuevas en la psicología.

Una de las primeras demostraciones de este fenómeno misterioso sucedió a principios de los 1930s. Joseph Banks Rhine, un psicólogo en Duke, había desarrollado un interés en la posibilidad de percepción extrasensorial, o E.S.P. Rhine diseñó un experimento en el que se usaban las cartas Zener, una baraja especial de veinticinco cartas impresas con

uno de cinco símbolos diferentes: se tomaba una carta de la baraja y al sujeto se le pedía que adivinara el símbolo. La mayoría de los sujetos de Rhine adivinaban correctamente alrededor del 20 por ciento de las cartas, como usted lo esperaba, pero un universitario de nombre Adam Linzmayer promedió casi 50 por ciento durante sus sesiones iniciales, y consiguió muchas rachas asombrosas, tales como adivinar nueve cartas en fila. Las probabilidades de que esto suceda por casualidad son casi de una en dos millones. Linzmayer lo hizo tres veces.

Rhine documentó tres resultados sensacionales en su cuaderno y preparó muchos escritos para ser publicados. Pero entonces, justo cuando comenzaba a creer en la posibilidad de la percepción extrasensorial, el estudiante perdió su espeluznante talento. Entre 1931 y 1933, Linzmayer trató de adivinar la identidad de muchas otras miles de cartas, pero su tasa de éxito ahora estaba apenas por encima de la casualidad. Rhine se vio obligado a concluir en que la "habilidad de percepción extrasensorial [del estudiante] había desaparecido bajo una disminución notable." Y linzmayer no fue el único sujeto en experimentar tal caída: en casi todos los casos en los que Rhine y otros documentaron la E.S.P., el efecto disminuyó dramáticamente con el tiempo. Rhine le llamó a esta tendencia el "efecto decadencia."

Schooler estaba fascinado por las batallas experimentales de Rhine. Aquí estaba un científico que había documentado repetidamente la decadencia de sus datos; parecía tener un talento para encontrar resultados que luego se hacían añicos. En 2004, Schooler se embarcó en una imitación irónica de la investigación de Rhine: trató de replicar este fracaso de replicar. En homenaje a los intereses de Rhine, decidió hacer pruebas de un fenómeno parasicológico conocido como precognición. El experimento en sí era franco: le mostraba rápidamente un conjunto de imágenes a un individuo y le pedía que identificara cada uno. La mayor parte del tiempo, la respuesta era negativa – las imágenes se mostraban con demasiada rapidez como para registrarlas. Luego Schooler seleccionaba al azar la mitad de las imágenes para mostrarlas nuevamente. Lo que quería saber era si las imágenes que se presentaban por segunda vez tenían más probabilidades de ser identificadas que la primera vez que se mostraron. ¿Podría la exposición posterior haber influenciado de alguna manera los resultados iniciales? ¿Podía el efecto llegar a ser la causa?

La locura de la hipótesis era el punto: Schooler sabe que la precognición carece de una explicación científica. Pero no estaba probando poderes extrasensoriales; estaba probando el efecto decadencia. "Al principio los datos fueron asombrosos, justo como lo esperábamos," dice Schooler. "No podía creer la cantidad de precognición que estábamos encontrando. Pero entonces, mientras seguíamos probando sujetos, la magnitud del efecto – una medida estadística estándar – "se fue haciendo cada vez más pequeña." Los científicos eventualmente examinaron a más de dos mil universitarios. "Al final, nuestros resultados se miraban exactamente como los de Rhine," dijo Schooler. "Hallamos este fuerte efecto paranormal, pero desapareció frente a nosotros."

La explicación más probable para la disminución es una que es obvia: la regresión al promedio. Es decir, a medida que se repite el experimento se cancela una primera casualidad estadística. Los poderes extrasensoriales de los sujetos de Schooler no declinaron – eran simplemente una ilusión que se esfumaba con el tiempo. Y no obstante Schooler ha notado que muchos de los datos que terminan disminuyendo parecen estadísticamente sólidos – es decir, contienen suficientes datos, tantos que cualquier regresión al promedio no debiese ser dramático. “Estos son los resultados que pasan todas las pruebas,” dice. “Las probabilidades de que sean algo aleatorio son típicamente bastante remotas, como una en un millón. Esto significa que el efecto de decadencia no debiese suceder casi nunca. ¡Pero pasa todo el tiempo! Caray, me ha pasado en múltiples ocasiones.” Y esta es la razón por la cual Schooler cree que el efecto decadencia merece más atención: su ubiquidad parece violar las leyes de la estadística. “Cada vez que comienzo a hablar de esto, los científicos se ponen muy nerviosos,” dice. “Pero todavía quiero saber qué les pasó a mis resultados. Como la mayoría de los científicos, asumí que se haría más fácil documentar mi efecto a lo largo del tiempo. Mejoré mucho con respecto a la realización de los experimentos, concentrándome en las condiciones que producen oscurecimiento verbal. Así que, ¿por qué sucedía lo opuesto? Estoy convencido de que podemos utilizar las herramientas de la ciencia para comprender esto. Aunque primero tenemos que admitir que tenemos un problema.”

En 1991, el zoólogo danés Anders Møller, de la Universidad de Uppsala en Suecia, hizo un sorprendente descubrimiento sobre el sexo, las golondrinas y la simetría. Se había sabido desde hacía mucho que la apariencia simétrica de una criatura estaba directamente asociada con la cantidad de mutaciones en su genoma, de modo que más mutaciones conducían a más “asimetrías fluctuantes.” (Una manera fácil de medir la asimetría en los humanos es comparar la extensión de los dedos de cada mano.) Lo que Møller descubrió es que era mucho más probable que las golondrinas hembras se aparearan con aves machos que tenían plumas largas y simétricas. Esto sugería que las quisquillosas hembras estaban usando la simetría como un poder para distinguir la calidad de los genes de los machos. El escrito de Møller, que fue publicado en la revista ***Nature***, muestra una cantidad abrumadora de investigación. Aquí estaba un indicador fácilmente medido y ampliamente aplicado de calidad genética, y se podía mostrar que las hembras gravitaban a su alrededor. La estética en realidad tenía que ver con la genética.

En los tres años que siguieron hubo diez pruebas independientes del papel de la asimetría fluctuante en la selección sexual, y nueve de ellas encontraron una relación entre la simetría y el éxito reproductivo de los machos. No importaba si los científicos estaban mirando los vellos de la mosca de la fruta o replicando los estudios de las golondrinas – las hembras parecían preferir a los machos cuyas mitades parecían reflejadas en un espejo. Antes de mucho, la teoría fue aplicada a los humanos. Por ejemplo, los investigadores encontraron que las mujeres preferían el olor de hombres simétricos, pero sólo durante la fase fértil del ciclo menstrual. Otros estudios afirmaban

que las hembras tenían más orgasmos cuando sus compañeros eran simétricos, mientras que un estudio por parte de antropólogos en Rutgers analizó cuarenta rutinas de danza provenientes de Jamaica y descubrió que los hombres simétricos eran catalogados de forma consistente como mejores bailarines.

Entonces la teoría comenzó a hacerse añicos. En 1994, hubo catorce pruebas publicadas de simetría y selección sexual, y solamente ocho encontraron una correlación. En 1995 hubo ocho trabajos sobre el tema, y solamente cuatro obtuvieron un resultado positivo. En 1998, cuando hubo doce investigaciones adicionales de asimetría fluctuante, solamente un tercio de ellas confirmó la teoría. Peor aún, incluso los estudios que mostraban resultado positivo también mostraban un continuo descenso en la magnitud del efecto. Entre 1992 y 1997, la magnitud del tamaño del efecto disminuyó en un ocho por ciento.

Y no es sólo la fluctuación de la asimetría. En 2001, Michael Jennions, un biólogo en la Universidad Nacional Australiana, se dedicó a analizar las "tendencias temporales" en un amplio rango de materias en ecología y biología evolutiva. Revisó cientos de trabajos y cuarenta y cuatro meta-análisis (es decir, síntesis estadísticas de estudios relacionados), y descubrió una disminución consistente del efecto a lo largo del tiempo, pues muchas de las teorías parecían desvanecerse en la irrelevancia. De hecho, aún cuando numerosas variables eran controladas – Jennions sabía, por ejemplo, que el mismo autor podía publicar varios trabajos críticos, lo que distorsionaría su análisis – aún había una disminución significativa en la validez de las hipótesis, a menudo en el lapso de tiempo de un año de publicación. Jennions admite que estos descubrimientos son perturbadores, pero expresa una renuencia a hablar de ellos públicamente. "Este es un asunto muy sensible para los científicos," dice. "Ya lo sabes, se supone que estamos tratando con hechos puros, el tipo de cosas que se supone deben resistir la prueba del tiempo. Pero cuando ves estas tendencias te vuelves un poco más escéptico de las cosas."

¿Qué pasó? Leigh Simmons, un biólogo de la Universidad de Australia Occidental, sugirió una explicación cuando me contó de su entusiasmo inicial por la teoría: "Estaba realmente emocionado por la asimetría fluctuante. Los primeros estudios hicieron que el efecto se viera muy robusto." Él decidió conducir algunos experimentos por cuenta propia, investigando la simetría en los escarabajos cornudos macho. "Desdichadamente, no pude encontrar el efecto," dijo. "Pero la peor parte fue que cuando presenté estos resultados nulos tuve dificultades en lograr que los publicaran. Las revistas sólo querían información que confirmara. Era una idea demasiado emocionante de refutar, al menos en ese entonces." Para Simmons, la abrupta subida y la lenta caída de la asimetría fluctuante es un claro ejemplo de un paradigma científico, una de esas modas pasajeras intelectuales que guían y constriñen el proceso de investigación: después que se propone un nuevo paradigma, el proceso de revisión por parte de iguales se ve inclinado hacia los resultados positivos. Pero luego, después de pocos años, los incentivos académicos experimentan un

giro – el paradigma se ha afianzado – de modo que los resultados más notables ahora son aquellos que desacreditan la teoría.

De igual manera, Jennions argumenta que el efecto decadencia es mayormente un producto de la parcialidad en cuanto a las publicaciones, o la tendencia de los científicos y de las publicaciones científicas a preferir información positiva por encima de los resultados nulos, que es lo que sucede cuando no se encuentra ningún efecto. Esta parcialidad fue identificada primero por el estadístico Theodore Sterling, en 1959, después que notó que noventa y siete por ciento de todos los estudios psicológicos publicados con datos estadísticamente significativos descubrieron el efecto que estaban buscando. Un resultado "significativo" se define como cualquier información puntual que sería producida por casualidad menos del cinco por ciento del tiempo. Esta prueba ubicua fue inventada en 1922 por el matemático inglés Ronald Fisher, quien escogió el cinco por ciento como la línea divisoria, algo arbitrariamente, porque facilitaba los cálculos con el lápiz y la regla de cálculo. Sterling miró que si el noventa y siete por ciento de los estudios de psicología estaban probando su hipótesis, o los psicólogos eran extraordinariamente suertudos o solamente publicaban los resultados de los experimentos exitosos. En años recientes, la parcialidad en las publicaciones se ha visto mayormente como un problema para los juicios clínicos, dado que las compañías farmacéuticas están menos interesadas en publicar resultados que no son favorables. Pero cada vez se hace más claro que la parcialidad en las publicaciones también produce importantes distorsiones en los campos de estudio sin grandes incentivos de las corporaciones, tales como la psicología y la ecología.

Aunque la parcialidad en las publicaciones juega casi con seguridad un papel en el efecto decadencia, sigue siendo una explicación incompleta. Por una cosa, fracasa al no brindar una explicación para el predominio inicial de resultados positivos entre los estudios que jamás son enviados siquiera a las publicaciones. También deja sin explicar la experiencia de personas como Schooler, que han sido incapaces de replicar sus datos iniciales a pesar de sus mejores esfuerzos. Richard Palmer, un biólogo en la Universidad de Alberta, quien ha estudiado los problemas que rodean a las fluctuaciones de la asimetría, sospecha que un asunto igualmente significativo es el reportaje selectivo de resultados – los datos que los científicos escogieron documentar para comenzar. La evidencia más convincente de Palmer descansa en una herramienta estadística conocida como gráfico de embudo. Cuando se ha llevado a cabo una gran cantidad de estudios sobre un solo tema, los datos debiesen seguir un patrón: los estudios con una gran medida de muestras debiesen todos agruparse alrededor de un valor común – el verdadero resultado – mientras que aquellos con una medida más pequeña de muestras deben exhibir una dispersión aleatoria dado que están sujetos a un mayor error de muestreo. Este patrón le da al gráfico su nombre, dado que la distribución se parece a un embudo.

El gráfico de embudo captura visualmente las distorsiones del reportaje selectivo. Por ejemplo, después que Palmer hubo trazado todos los estudios sobre la asimetría

fluctuante notó que la distribución de resultados con las muestras de tamaño más pequeño no eran aleatorias en lo absoluto sino que, en vez de eso, se inclinaban de manera sesgada hacia resultados positivos. Desde entonces Palmer ha documentado un problema similar en muchas otras áreas de estudio impugnadas. “Una vez que me di cuenta que el reportaje selectivo está en todas partes en la ciencia, me deprimí bastante,” me dijo Palmer. “Como investigador estás siempre consciente de que podría haber algunos patrones no aleatorios, pero no tenía idea de cuán generalizado es esto.” En una reseña reciente Palmer resumió el impacto del reportaje selectivo en su campo: “No podemos escapar de la perturbadora conclusión de que algunas – quizá muchas – muy queridas generalidades han sido, en el mejor de los casos, exageradas en su significación biológica, y en el peor de ellos, son una ilusión colectiva nutridas por unas creencias a priori a menudo repetidas.”

Palmer enfatiza que el reportaje selectivo no es lo mismo que el fraude científico. Más bien, el problema parece ser un problema de omisiones sutiles y percepciones erróneas inconscientes a medida que los investigadores batallan para darle sentido a sus resultados. Stephen Jay Gould se refirió a esto como el proceso “de la horma de zapato.” “Mucho de la medición científica es algo realmente fuerte,” me dijo Simmons. “Si estás hablando de la fluctuación en la simetría, entonces es asunto de diferencias minúsculas entre los lados derecho e izquierdo de un animal. Son milímetros de una pluma de la cola. Y así, puede ser que un investigador sepa que está midiendo a un buen macho” – un animal que se ha apareado exitosamente – “y sabe que se supone que sea simétrico. Bueno, ese acto de medición va a ser vulnerable de toda clase de parcialidades de percepción. Esa no es una declaración cínica. Esa es simplemente la manera en que trabajan los seres humanos.”

Uno de los ejemplos clásicos de reportaje selectivo concierne a la prueba de la acupuntura en diferentes países. Aunque la acupuntura es ampliamente aceptada como un tratamiento médico en varios países de Asia, su uso es mucho más impugnado en Occidente. Estas diferencias culturales han influenciado profundamente los resultados de las pruebas clínicas. Entre 1966 y 1995 hubo cuarenta y siete estudios de acupuntura en China, Taiwán y Japón, y cada una de las pruebas concluyó en que la acupuntura era un tratamiento efectivo. Durante el mismo período hubo noventa y cuatro pruebas clínicas de acupuntura en los Estados Unidos, Suiza y el Reino Unido, y solamente cincuenta y seis por ciento de estos estudios encontraron beneficios terapéuticos. Como señala Palmer, esta amplia discrepancia sugiere que los científicos encuentran maneras para confirmar sus hipótesis preferidas ignorando lo que no quieren ver. Nuestras creencias son una forma de ceguera.

John Ioannidis, un epidemiólogo en la Universidad de Stanford, argumenta que tales distorsiones son un asunto serio en la investigación biomédica. “Estas exageraciones son la razón por la cual la decadencia se ha vuelto tan común,” dice. “Sería en realidad grandioso si los estudios iniciales nos dieran un resumen preciso de las cosas. Pero no lo hacen. Y así lo que sucede es que desperdiciamos una gran cantidad de dinero tratando a millones de pacientes y haciendo una gran cantidad de estudios de seguimiento en otros

temas basados en resultados que son engañosos." En 2005, Ioannidis publicó un artículo en el *Journal of the American Medical Association* que daba una mirada a los cuarenta y nueve estudios más citados de investigación clínica en tres grandes publicaciones médicas. Cuarenta y cinco de estos estudios reportaban resultados positivos sugiriendo que la intervención que estaba siendo probada era efectiva. Debido a que la mayoría de estos estudios eran pruebas aleatorias de control – el "patrón oro" de la evidencia médica – tenían la tendencia a tener un impacto significativo en la práctica clínica, y conducían a la propagación de tratamientos tales como la terapia de reemplazo de hormonas para las mujeres menopáusicas y bajas dosis de aspirinas diariamente para prevenir los ataques al corazón y los derrames cerebrales. Sin embargo, los datos que Ioannidis encontró eran perturbadores: de las treinta y cuatro pruebas que habían sido sujetas a replicación, cuarenta y un por ciento o habían sido directamente contradichas o la proporción de sus efectos se había reducido significativamente.

La situación es incluso peor cuando un tema está de moda. Por ejemplo, en años recientes ha habido cientos de estudios acerca de los varios genes que controlan las diferencias en el riesgo de las enfermedades entre hombres y mujeres. Estos descubrimientos han incluido todo desde las mutaciones responsables por el riesgo incrementado de la esquizofrenia hasta los genes que subyacen a la hipertensión. Ioannidis y sus colegas estudiaron cuatrocientos treinta y dos de estas afirmaciones. Rápidamente descubrieron que la amplia mayoría tenía serios defectos. Pero el hecho más perturbador surgió cuando miró la prueba de replicación: de cuatrocientas treinta y dos afirmaciones solamente una era consistentemente replicable. "Esto no significa que ninguna de estas afirmaciones resultará siendo verdadera," dice. "Pero, dado que la mayoría de ellas ha sido hecha de manera defectuosa, yo no aguantaría la respiración."

Según Ioannidis, el principal problema es que demasiados investigadores se involucran en lo que llama "persecución de significado," o encontrar maneras de interpretar los datos de modo que pasen la prueba estadística de significado – la frontera del noventa y cinco por ciento inventada por Ronald Fisher. "Los científicos están tan impacientes por pasar esta prueba mágica que comienzan a jugar con los números, tratando de encontrar algo que parezca digno," dice Ioannidis. En años recientes, Ioannidis se ha tornada cada vez más franco con respecto al carácter omnipresente del problema. Uno de sus trabajos más citados tiene un título deliberadamente provocativo: "Por Qué la Mayor Parte de las Investigaciones Publicadas Es Falsa."

El problema del reportaje selectivo se arraiga en un defecto cognitivo fundamental, el cual es que nos gusta probarnos a nosotros mismos que tenemos la razón y odiamos estar equivocados. "Se siente una gran sensación cuando se valida una hipótesis," dijo Ioannidis. "Se siente aún mejor cuando tienes un interés financiero en la idea o si tu carrera depende de ella. Y esa es la razón por la cual todavía puede verse, aún después que una afirmación ha sido sistemáticamente refutada" – cita, por ejemplo, los primeros trabajos sobre la terapia de reemplazo de hormonas, o las afirmaciones con respecto a

varias vitaminas – “a algunos investigadores tercos citar los primeros pocos estudios que muestran un fuerte efecto. Ellos realmente quieren creer que es verdad.”

Esa es la razón por la cual Schooler argumenta que los científicos necesitan ser más rigurosos con respecto a la recolección de datos antes de publicar. “Estamos desperdiciando demasiado tiempo corriendo detrás de malos estudios y experimentos con poca potencia,” dice. La actual “obsesión” con la replicabilidad funciona como distracción del verdadero problema, el cual es el diseño defectuoso. Señala que nadie ni siquiera trata de replicar la mayoría de trabajos de ciencia – simplemente hay demasiados. (De acuerdo a *Nature*, una tercera parte de todos los estudios jamás son siquiera citados, mucho menos repetidos.) “He aprendido de la manera dolorosa a ser excesivamente cuidadoso,” dice Schooler. “Todo investigador debiese tener que explicar en detalle, por adelantado, cuántos sujetos van a usar, y qué exactamente están probando, y qué constituye un nivel suficiente de prueba. Tenemos las herramientas para ser mucho más transparentes con respecto a nuestros experimentos.”

En un trabajo aún por aparecer Schooler recomienda el establecimiento de una base de datos de código abierto, y en base a él a los investigadores se les requiere que bosquejen sus investigaciones planeadas y documenten todos sus resultados. “Pienso que esto proveería un enorme incremento en el acceso al trabajo científico y nos brinda una mejor manera de juzgar la calidad de un experimento,” dice Schooler. “Nos ayudaría a tratar finalmente con todos estos asuntos que el efecto decadencia está exponiendo.”

Aunque tales reformas mitigarían los peligros de la parcialidad en la publicación y el reportaje selectivo, aún no borrarían el efecto decadencia. Esto en gran parte es así porque la investigación científica será siempre ensombrecida por una fuerza que no puede ser frenada, solamente contenida: la pura aleatoriedad. Aunque se ha hecho poca investigación sobre los peligros experimentales del azar y la casualidad, la investigación que existe no es alentadora.

A fines de los años noventa, John Crabbe, un neurocientífico de la Universidad de la Ciencia y la Salud de Oregon, condujo un experimento que mostró cómo eventos desconocidos de casualidad pueden sesgar las pruebas de replicabilidad. Él realizó una serie de experimentos sobre la conducta de los ratones en tres diferentes laboratorios de ciencia: en Albany, New York; Edmonton, Alberta; y Portland, Oregon. Antes de conducir el experimento, trató de estandarizar todas las variables sobre las cuales pudiera pensar. Se crió la misma clase de ratones en el mismo tipo de recinto, con la misma marca de capa de serrín. Habían sido expuestos a la misma cantidad de luz incandescente, vivían con la misma cantidad de compañeros y se les alimentaba con las mismas bolitas de comida. Cuando se trabajaba con los ratones se usó el mismo tipo de guante quirúrgico, y cuando fueron probados fue con el mismo equipo, al mismo tiempo por la mañana.

La premisa de esta prueba de replicabilidad, claro está, es que los mismos laboratorios debían haber generado el mismo patrón de resultados: "Si algún grupo de experimentos debió haber pasado tendría que haber sido el nuestro," dice Crabbe. "Pero no fue así como resultó." En un experimento, Crabbe le inyectó cocaína a una variedad particular de ratón. En Portland los ratones a los cuales se les administró la droga se movieron, en promedio, seiscientos centímetros más de lo que normalmente lo hacían; en Albany se movieron setecientos un centímetros adicionales. Pero en el laboratorio de Edmonton se movieron más de cinco mil centímetros adicionales. Se observaron desviaciones similares en una prueba de ansiedad. Además, estas inconsistencias no seguían algún patrón detectable. En Portland, una variedad de ratón probó ser la más ansiosa, mientras que en Albany otra variedad ganó esa distinción.

La inquietante implicación del estudio de Crabbe es que una buena cantidad de datos científicos extraordinarios no son más que ruido. La hiperactividad de aquellos ratones cocainómanos de Edmonton no era un nuevo hecho interesante – era un resultado carente de significado, un subproducto de variables invisibles que no entendemos. El problema, claro está, es que tales descubrimientos tan dramáticos también son los que tienen más probabilidades de ser publicados en diarios de prestigio, puesto que los datos son estadísticamente significativos y totalmente inesperados. Se firman las subvenciones y se conducen estudios de seguimiento. El resultado final es un accidente científico que puede requerir de años para ser aclarado.

Esto sugiere que el efecto decadencia es en realidad una decadencia de la ilusión. Mientras Karl Popper imaginaba que la falsificación se llevaba a cabo con un solo y definitivo experimento – Galileo refutó la mecánica aristotélica en una tarde – el proceso resulta ser mucho más complicado que eso. Muchas teorías científicas siguen siendo consideradas verdaderas incluso después de fracasar en numerosas pruebas experimentales. El oscurecimiento verbal podría exhibir el efecto decadencia, pero se sigue confiando extensamente en él dentro del campo. Lo mismo puede decirse de una gran cantidad de fenómenos, desde los desaparecidos beneficios de los anti sicóticos de segunda generación hasta la débil proporción de acoplamiento exhibida por los neutrones en deterioro, que parece haber caído por más de diez estándares de desviación entre 1969 y 2001. Incluso la ley de la gravedad no siempre ha sido perfecta en predecir fenómenos del mundo real. (En una prueba, los físicos que medían la gravedad por medio de profundas perforaciones en el desierto de Nevada encontraron una discrepancia de 2.5% entre las predicciones teóricas y los datos recopilados.) A pesar de estos descubrimientos, los anti sicóticos de segunda generación todavía son ampliamente recetados, y nuestro modelo del neutrón no ha cambiado. La ley de la gravedad sigue siendo la misma.

Tales anomalías demuestran el carácter escurridizo del empirismo. Aunque muchas ideas científicas generan resultados en conflicto y sufren de defectos en la magnitud de los efectos, siguen siendo citadas en los libros de texto y dirigen la práctica médica estándar. ¿Por qué? Porque estas ideas parecen verdaderas. Porque tienen sentido. Porque

no podemos soportar que se vayan. Y esta es la razón por la cual el efecto decadencia es tan inquietante. No porque revela la falibilidad humana de la ciencia, en la cual se tuercen los datos y las creencias moldean las percepciones. (Tales deficiencias no son una sorpresa, al menos para los científicos.) Y no porque revela que muchas de nuestras más emocionantes teorías son modas pasajeras que pronto serán rechazadas. (Esa idea ha andado por ahí desde Thomas Kuhn.) El efecto decadencia es inquietante porque nos recuerda cuán difícil es probar cualquier cosa. Nos gusta pretender que nuestros experimentos definen la verdad para nosotros. Pero con frecuencia ese no es el caso. Sólo porque una idea es verdadera no significa que puede ser probada. Y sólo porque una idea puede ser probada no significa que es verdadera. Cuando se han hecho los experimentos, todavía tenemos que escoger qué vamos a creer. ♦

Este artículo fue originalmente publicado en inglés por *The New Yorker*. Puede ubicar el artículo original en la dirección:
http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer?currentPage=1

Traducción de Donald Herrera Terán, para www.contra-mundum.org