ID: 10.33881/2027-1786.RIP.14109

ĬbērAM
Investigación | Research

Programas de intervención y replicabilidad

consideraciones sobre su evaluación en Psicología

Intervention and replicability programs: considerations on their Evaluation of in Psychology

José Fernando **Mora-Romo**

Photo By/Foto: Kheng Guan Toh

Rip 14¹
Volumen 14 #1 ene-abr
14 Años

Revista Iberoamericana de
Psicología
ISSN-l: 2027-1786 | e-ISSN: 2500-6517
Publicación Cuatrimestral

Revista Iberoamericana de
Psicología
ISSN-l: 2027-1786 | e-ISSN: 2500-6517
Publicación Cuatrimestral

ID: 10.33881/2027-1786.RIP.14109

Title: Intervention and replicability programs

Subtitle: Considerations on their evaluation of in psychology

Título: Programas de intervención y replicabilidad

Subtítulo: Consideraciones sobre su evaluación en psicología

Alt Title / Título alternativo:

[en]: Intervention and replicability programs: Considerations on their Evaluation of in Psychology

[es]: Programas de intervención y replicabilidad: Consideraciones sobre su Evaluación en Psicología

Author (s) / Autor (es):

Mora-Romo

Keywords / Palabras Clave:

[en]: Intervention; Evaluation; Replication; Bayes' Theorem; Size Effect; Confidence Intervals

[es]: Intervención; Evaluación; Replicación; Teorema de Bayes; Tamaño del Efecto; Intervalos de Confianza

Submited: 2020-06-08

Acepted: 2020-08-02

Resumen

En este trabajo se hace un repaso acerca de qué es la evaluación de programas de intervención, así como el procedimiento y finalidad que se ha buscado brindarle desde la psicología; y el procedimiento utilizado para llevar a cabo la replicación de estos estudios, especialmente en psicología, para así poder dar contexto a la llamada "crisis de confianza" en psicología. Esto con la finalidad de proponer que la replicabilidad de programas de intervención no podría lograrse de forma satisfactoriamente sin antes llevar a cabo un proceso de evaluación del estudio original. Para esto, se proponen tres parámetros, Factor de Bayes, Tamaño del efecto e Intervalos de Confianza; que han mostrado su utilidad en la literatura de habla inglesa, pero, sin embargo, no se ha difundido de manera extensa en aquella de habla hispana. Se propone lograr una mayor difusión de la estadística Bayesiana para solventar las problemáticas ocasionadas por la estadística frecuentista de prueba de significancia de hipótesis nula, puesto que una de las finalidades del teorema de Bayes es la acumulación y actualización continua del conocimiento obtenido de replicaciones, sin dejar de lado la evaluación de los grados de certeza de los resultados obtenidos. Así, este trabajo está dirigido para profesionistas nuevos y veteranos en el ámbito de la evaluación, que empiezan a adentrarse en la estadística como una forma de evaluar programas de investigación e intervención, por lo que se ha buscado explicar los parámetros propuestos de la forma más clara y concisa posible

Abstract

In this work a review is made about what evaluation of intervention programs is, as well as the its procedure and purpose that it has been sought to provide from psychology; and the procedure used to carry out the replication of these studies, especially in psychology, in order to give context to the so-called "crisis of confidence" in psychology. This in order to propose that replicability of intervention programs could not be satisfactorily achieved without first carrying out an evaluation of the original study. For this, three parameters are proposed, Bayes Factor, Effect Size and Confidence Intervals; which have shown their usefulness in English-speaking literature, but, nevertheless, it has not been widely used in Spanish-speaking literature. It is proposed to achieve a wider dissemination of Bayesian statistics to solve the problematics caused by the frequentist statistics of the Null Hypothesis Significance Test, since one of the purposes of Bayes' theorem is the accumulation and continuous updating of the knowledge obtained from replications, without leave aside the evaluation of the degrees of certainty of the results obtained. Thus, this work is aimed at new and veteran professionals in the field of evaluation, who are beginning to delve into statistics as a way of evaluate research and intervention programs, which is why we have sought to explain the parameters proposed in the most clear and concise way as possible

Citar como:

Mora-Romo, J. F. (2021). Programas de intervención y replicabilidad: Consideraciones sobre su evaluación en psicología. Revista Iberoamericana de Psicología , 14 (1), 93-104. Obtenido de: https://reviberopsicologia.ibero.edu.co/article/view/1867

José Fernando **Mora-Romo**, Psi
ORCID: *https://orcid.org/0000-0002-6201-4622*

Source | Filiacion:
Egresado de Universidad Autónoma de Zacatecas

BIO:
Investigador

City | Ciudad:
Zacatecas [mx]

e-mail:
j_fmora@hotmail.com

Revista Iberoamericana de **Psicología**
ISSN-L: 2027-1786

Programas de intervención y replicabilidad

consideraciones sobre su evaluación en Psicología

Intervention and replicability programs: considerations on their Evaluation of in Psychology

José Fernando **Mora-Romo**

Sabemos que un programa puede obtener resultados favorables en las pruebas de significación estadística, pero no conocemos con certeza qué efecto tiene realmente el programa, o qué grado de certeza se le da a la hipótesis alternativa. De saber este dato, la toma de decisiones en la evaluación de programas de intervención en psicología podría beneficiarse puesto que pudiéramos escoger entre un programa cuyos resultados fueron *p <.001*, un factor bayes *(K) <5 y d: .2;* o un programa cuyos resultados fueron *p <.05* pero con *K >13 y d: .7.* Esto debido a que la información que el p-valor nos está ofreciendo por sí solo es que de la muestra que participaron en el programa, hay un 5% de probabilidad de que los resultados obtenidos entre grupos sean debido al azar **(Dahiru, 2008**); no nos está diciendo que, en futuras replicaciones, esa misma probabilidad se mantenga, puesto que no contempla la acumulación, y actualización en consecuencia, de datos **(Hubbard y Lindsay, 2008).**

Esta confusión sobre los alcances y limitaciones prácticos de estas herramientas estadísticas de evaluación nos pudiera dar una idea de que la "crisis de la replicabilidad" en ciencias sociales es algo que debimos de haber esperado.

Este trabajo tiene tres objetivos: hacer una revisión respecto qué es la evaluación y su desarrollo desde la psicología; revisar qué es la replicabilidad en general, y en psicología en específico, y la "crisis" entorno a ella; y, por último, plantear la utilidad de tres parámetros para la interpretación de los resultados de programas de intervención y replicación.

Evaluación de programas de intervención

¿Qué es la evaluación?

Una evaluación, según la World Health Organization **(2013),** es una valoración de los resultados, procesos, factores contextuales y causales llevados a cabo para lograr un entendimiento de los logros obtenidos a través de un programa de intervención. Para esto, la National Center for Chronic Disease Prevention and Health Promotion, **(2011),** plantea que la elaboración de un plan de evaluación es esencial para esto, donde describa el programa completo y sus actividades, así como la forma en que estas últimas se orientan para alcanzar los objetivos del primero.

Fernández-Ballesteros **(2013a),** habla de dos procesos distintos dentro de la evaluación. Uno descriptivo-predictivo para conocer las condiciones de causa y mantenimiento de una situación dada; y otro interventivo-valorativo, en el cual se incluye el plan y valoración de tratamiento; diseño, administración y evaluación continua, y la valoración final de todo el plan de evaluación. Por su parte, Ortegón, Pacheco y Prieto **(2015),** argumentan que la evaluación puede comenzar desde antes de llevar a cabo la intervención (durante el diseño del programa), hasta incluso alcanzar "varios años después de completada la ejecución, en el caso de evaluaciones de impacto y/o sustentabilidad" **(p. 48).**

Shipman **(1989)** ofrece una propuesta sobre cómo orientar este trabajo cuando se consideran a poblaciones diferentes, aun teniendo los mismos, o diferentes, objetivos, estableciendo un componente descriptivo enfocado a la problemática planteada y otro a la valoración de la necesidad del programa, su implementación y los efectos logrados. Otra problemática, como argumentan Ludwig, Kling y Mullainathan**. (2011),** es poder generalizar los resultados obtenidos en un programa a escalas comunitarias, estatales o federales. Para lograr esto, los autores proponen la importancia de conocer los mecanismos conductuales, "mediadores" **(p. 21)**, a través de experimentos de mecanismos (Mechanism experiments, en inglés), los cuales ponen a prueba la cadena causal entre los mediadores y los resultados obtenidos. En otras palabras, los experimentos de mecanismos se referirían a lo que Bunge **(2012, p. 338**) llama modelo de caja dinámica puesto que considera que la variable interviniente, en este caso los mediadores, representa el mecanismo que transforma las entradas en salidas, es decir, en los resultados obtenidos. Considerando esto, la finalidad de la evaluación, como argumenta Gairín **(2010),** supone una comprensión de la realidad estudiada de manera sistemática y rigurosa, regida por principios de utilidad y participación, atendiendo al contexto de la población y a efectos secundarios provocados, por lo que debe ser utilizada éticamente.

La implementación de la evaluación de programas ha contribuido al aumento del conocimiento sobre el abordaje de las problemáticas sociales, sin embargo, siguiendo a Rossi **(1987)** este conocimiento abarca desde un punto de vista optimista hasta uno pesimista, pesimismo que afloraría en la llamada "crisis de la replicabilidad" que abordaremos más adelante. Por ahora pasaremos a considerar la evaluación de programas desde la psicología.

La evaluación desde la psicología

Al escribir "la evaluación desde la psicología", y no en la psicología, se busca llevar a cabo la reflexión sobre que, si bien el proceso es similar a lo descrito en la subsección anterior, la evaluación que parte de la psicología lo hace con modelos propios. En este sentido, con base en el desarrollo del trabajo evaluativo, Fernández-Ballesteros **(2013b)** presenta seis modelos que han servido de base para los programas de evaluación en psicología: Atributivo, dinámico, médico, conductual, cognitivo y constructivista **(p. 36).** Dichos modelos hacen énfasis en factores endógenos o exógenos al explicar los resultados obtenidos. Otra reflexión que se busca hacer es que la evaluación debe ser conceptualizada como una estrategia para iniciar una acción de cambio **(Cook, 2014),** identificando el grado en que las necesidades planteadas distan de las obtenidas por el programa de intervención.

Este distanciamiento se ve reflejado en la metáfora ecológica **(Kelly, 1966),** que propone patrones de comportamiento caracterizados por la situacionalidad social. Por ello, una conceptualización ecológica puede orientar los planes de evaluación hacia la toma de decisiones respecto a qué acciones tomar para promover el bienestar social **(Miller, 2014),** adaptándolos a la dinámica social. Si bien, considerar la evaluación como impulsor del cambio social es una propuesta desde la psicología comunitaria **(Sheldon y Wolfe, 2014),** debido a su inclusión en la Asociación Americana de Evaluación **(Newman et al., 1995),** prueba la existencia de un entendimiento recíproco al orientar el cambio social con base a los resultados obtenidos en las evaluaciones de forma competente.

Otro esfuerzo que se ha realizado desde la psicología para los programas de evaluación es la Psicología Basada en Evidencia (PBE), para establecer la eficacia de las que, quizás en la actualidad, ya superan los 400 tipos de terapias reportadas por Kazdin **(1986).** La PBE basa sus evaluaciones **(Ybarra et al., 2015)** con base en la eficacia, la generalización y la viabilidad, y la relación costo-beneficios de la intervención; integrando evidencia científica con la experiencia clínica, además de considerar el contexto, cultura y preferencia del participante. Moriana y Martínez **(2011)** reportan dos instituciones relevantes en la PBE: La American Psychological Association yla National Institute for Health and Clinical Excellence. Estas instituciones evalúan los procedimientos de los tratamientos, así como depurando aquellos que no contribuyan a la mejora de la atención eficaz y eficiente, ofreciéndolos siempre como orientaciones y no como normativas de acción. El tema de la PBE ha sido abordado por parte de asociaciones e investigadores quienes han aportado a su desarrollo **(American Psychological Association, 2006; Echeburúa, De Corral y Salaberría., 2010; Hunsley, 2017; Kazdin, 2008).**

Ahora, podemos argumentar que la evaluación de programas valora la eficacia-eficiencia, al igual que la valoración de la predicción de dichos resultados **(Martorell y Gómez, 2010)**. Al hacer énfasis en la descripción de los diseños y objetivos de programas de evaluación, no debemos olvidar que quienes nos brindan la información para llevar esto a cabo son personas. Así, la excesiva ansiedad por evaluación **(Donaldson, Gooler y Scriven, 2002)** por una percepción de poca sistematicidad en la evaluación por parte de los participantes, o malas experiencias con evaluadores en el pasado, influye en los resultados; ocasionando un sesgo en los resultados que provocarían problemáticas como las argumentadas en la siguiente sección.

Con lo anterior mencionado, se pretende que se considere a la evaluación de programas en general, y de psicología en particular, como un requisito antes de considerar llevar a cabo una replicación, puesto que nos puede proporcionar información valiosa sobre lo que se pudiera esperar encontrar en futuros trabajos.

Crisis de la replicabilidad en psicología

¿Qué es la replicabilidad?

La replicabilidad es definida como la capacidad de duplicar los mismos resultados de un estudio anterior si el mismo procedimiento es realizado, aun cuando nueva información es recolectada **(Bollen et al., 2015).** Esta es considerada como "reproducibilidad de resultados" **(Goodman, Fanelli y Ioannidis, 2016, p. 2),** la cual es evaluada a la luz de la evidencia acumulada resultante de diversas replicaciones. Esto puede fortalecer la veracidad de una teoría o un modelo de intervención **(Park, 2004)** ya sea que se considere a poblaciones similares (reproductibilidad) o a poblaciones diferentes (generalización). Se han argumentado tipos de replicación **(Tsang y Kwan, 1999)** considerando si se usan los mismos procedimientos o no; reconceptualizaciones sobre qué se concibe como una replicación lo suficientemente buena **("Good Enough Replication", Singh, Ang y Leong, 2003, p. 539**); y tipos y aspectos a considerar en el diseño de replicaciones **(Hendrick, 1990).**

Sin duda un aspecto importante en la replicación en investigación es su carácter acumulativo de conocimiento. Sin embargo, pudiéramos considerar dos grandes enfoques: Replicaciones cercanas, tradicionalmente llamadas replicaciones conceptuales, aquellas que asumen que una replicación absolutamente exacta es imposible de hacer **(Tsang y Kwan, 1999; Brandt et al., 2014)**; y las replicaciones directas, que asumen que, siguiendo los mismos procedimientos, y con un adecuado manejo estadístico, es posible obtener los mismos resultados **(Simons, 2014).** La primera es defendida por aquellos que optan por el realismo crítico como epistemología **(Bunge, 1983, p 93),** la cual afirma que podemos conocer la realidad, aunque no completamente; mientras que la segunda es defendida por aquellos que optan por un realismo ingenuo **(Bunge, 1993 p. 230)** el cual afirma que podemos conocer completamente la realidad a través de los sentidos. Las replicaciones conceptuales **(Crandall y Sherman, 2016),** proponen poner a prueba la hipótesis del trabajo original, pero considerando procedimientos, variables, poblaciones y diseños diferentes. Es decir, se estaría buscando replicar no un programa de intervención, sino los supuestos de una teoría en concreto, por ello este tipo de replicación es de ayuda a establecer consenso respecto al procedimiento inferencial de los resultados con base en su acumulación **(p. 95).**

Llegados a este punto, podemos considerar que las replicaciones son de gran utilidad para constatar el grado de exactitud de resultados anteriores. Sin embargo, hay una diferencia entre llevar a cabo replicaciones y "entender, apreciar y apoyarlas" **(Makel y Plucker, 2014a, p. 29).** Esto debido a la concepción respecto a que la única contribución hacia una disciplina científica es mediante la investigación novedosa, dejando de lado la replicación. Pero, ¿cómo saber si los resultados de una replicación son adecuados y pueden contribuir al desarrollo de la disciplina? Una manera es sortear los errores muestrales inherentes a todo estudio de replicación que ocasiona variación en los resultados **(Spence y Stanley, 2016).** Para esto es necesario establecer un rango en el cual sea aceptable dicha variación. Tradicionalmente, se emplean los intervalos de confianza, pero, como lo argumentan los autores **(p. 3),** tiene una problemática similar al p-valor, misma que fue comentada en la sección de introducción, el cual puede provocar una errónea interpretación de los resultados. Para aclarar dicho error, una explicación de los Intervalos de Confianza, y sus supuestos, será presentada en la sección siguiente, pero por ahora pasemos a revisar la concepción que se tiene sobre la replicación en la psicología.

Argumentaciones en torno a la replicabilidad en la psicología

De igual forma, en la psicología se aprecia la importancia que tienen las replicaciones como un procedimiento para validar los hallazgos empíricos. Este procedimiento se ha centrado **(Francis, 2012)** en rechazar la hipótesis nula, lo cual ha provocado el "sesgo de publicación" **(p. 585).** Esto es la exageración o supresión del tamaño del efecto, para sobrestimar la probabilidad de rechazar la hipótesis nula. Pero claro, si no se publican los resultados con poco, o nulo efecto, ¿cómo podremos calcular **(p. 588; p. 592)** el sesgo de estas replicaciones?

En este sentido, García-Garzón, Lecuona y Carbajal **(2018)** presentan una serie de recomendaciones para facilitar una psicología científica abierta por medio de pre-registros de proyectos científicos antes de comenzar la investigación **(p. 79)** con la finalidad de disminuir las prácticas cuestionables de investigación **(PCI) (p. 76)** que han obstaculizado el desarrollo de programas de replicación en la disciplina. Otras recomendaciones para llevar a cabo replicaciones en psicología fueron realizadas por Tackett et al. **(2017).** Entre estas se vuelven a mencionar la necesidad de disminuir la incidencia de PCI y los pre-registros, empezar a considerar seriamente el tamaño del efecto (aspecto que revisaremos más adelante), así como las replicaciones independientes e identificar los programas de investigación más susceptibles a presentar sesgos en la interpretación de resultados

Otro sentido que se le ha dado a la replicación en psicología es propuesta por Ordoñez **(2014)** quien lo plantea como un papel pedagógico orientado a estudiantes de pregrado, puesto que los vincula a la práctica investigativa para promover el razonamiento científico con base en el entendimiento conceptual y metodológico de la profesión para el desarrollo del conocimiento y resolución de problemas, por lo que el uso pedagógico de replicaciones es fundamental para lograr la relación entre el conocimiento de la teoría y la metodología. Este aspecto resulta importante para resaltar la importancia de las replicaciones dentro de la psicología, ya que, como mencionan Koole y Lakens, **(2012),** la integración de este tipo de procedimiento dentro de los programas educativos puede brindar la oportunidad de reconocer el valor dentro de la formación de futuros profesionales capaces de llevar a cabo trabajo de investigación de calidad considerando los cuidados metodológicos que deban implementar, así como al desempeño profesional al abrir un campo para la formación de grupos colaborativos en la revisión y elaboración de este tipo de procedimientos.

RIP 14 (1) pág. 93-104
Revista Iberoamericana de Psicología
ISSN-L: 2027-1786

A pesar de que se pueda tener una buena percepción sobre la utilidad de las replicaciones en la ciencia en general, y en psicología en específico, Makel, Plucker y Hegarty **(2012),** al revisar 100 revistas de publicaciones de psicología, encontraron que solamente un 1.57% de las publicaciones utilizan la palabra replicación, porcentaje que disminuye a 1.07% al momento en que revisaron 500 artículos de esas revistas para corroborar que, de hecho, se tratasen de replicaciones como tal. Sin embargo, los autores mencionan que esta práctica ha venido en aumento desde 1990, encontrando aún mayor incremento desde el 2010; por lo que se ha empezado a desarrollar una valoración hacia este tipo de investigación, que pudiéramos considerar como debido a una demanda social, la cual consideramos en la siguiente subsección.

Crisis de la replicabilidad

La crisis de la replicabilidad en psicología, parte de la consideración respecto que se han encontrado dificultades en replicar los resultados obtenidos. Esto debido a varios aspectos **(Makel y Plucker, 2014b)** como la falta de aleatorización y muestras pequeñas de participantes, además de la ocurrencia de las PCI, mencionadas anteriormente, como la inclusión de información hasta poder rechazar la hipótesis nula, no reportar las replicaciones fallidas o eliminar los casos extremos aleatorios.

Como se presentó en la introducción, el apoyo recurrente, y desconocimiento de las limitaciones y postulados, al p-valor, ha provocado una malinterpretación de los resultados obtenidos, especialmente en psicología **(Cohen, 1994, p. 997).** El autor argumenta que, siguiendo con una errónea interpretación del p-valor, se crea la creencia que, al rechazar la hipótesis nula en un estudio, es probable que la replicación del mismo obtendrá resultados similares. Esto se conoce como el "error de ilusión de la identificación bayesiana" **(Bayesian id's wishful thinking error, Cohen, 1994, p. 999);** y esta malinterpretación pudiera ser uno de los causantes de la crisis de la replicabilidad en psicología, especialmente cuando se emplean métodos cuantitativos.

Otra posible causante de esta crisis se argumenta que sea la potencia estadística **(Schmidt y Oh, 2016)** siendo que este suele tener un rango entre .40 a .50, el cual no ha aumentado desde los últimos 50 años **(p. 33).** Esta potencia estadística plantea que puede haber dificultades en replicar resultados si este es *<.80* **(Maxwel, Lau y Howard, 2015, p. 489)**, un tamaño de la muestra inapropiado y no considerar la variación entre el tamaño del efecto del estudio original con el de la replicación. También se lleva considerar **(Schmidt y Oh, 2016, p. 33)** que el error muestral ha sido infravalorado por los investigadores, definiéndolo como la diferencia entre los valores estimados de una población y el valor real que presentaría. Como consideración de lo anterior, y prosiguiendo con el tema, se llega a cuestionar si la falta de replicación de los resultados del estudio original se debe a la ocurrencia de falsos-positivos en los primeros resultados. Sin embargo, al igual que se comentaba al inicio de este trabajo, los procedimientos estadísticos son sólo un aspecto a considerar dentro de este tipo de trabajos, siendo la parte cualitativa, como pueden ser la falta de recomendaciones para establecer un diseño metodológico adecuado, por ejemplo, el propuesto por **(Hendrick, 1990),** otro aspecto a tomar en cuenta.

Sin embargo, argumentan **(Earp y Trafimow, 2015)** que los mayores problemas que plantea esta crisis es la falta de incentivos debido al poco prestigio de realizar replicaciones; y una poca claridad respecto a los alcances de las replicaciones, es decir ¿Qué se espera inferir después de que se realiza una replicación? ¿Que se logró corroborar/falsear de la teoría del estudio original, o que los resultados del estudio original eran erróneos? **(p. 2).** Estas preguntas se podrían responder al delimitar el diseño de replicaciones dentro de las replicaciones directas o conceptuales, comentadas anteriormente.

A pesar del cuidadoso manejo metodológico que se realice para fomentar la validez interna, y considerar la potencia estadística para cuidar la validez externa, hay otros aspectos que influyen en la falta de replicabilidad en psicología. En este sentido, Santos **(2013)** argumentaba sobre la influencia de variables moderadoras, los grados de libertad del investigador (PCI), eliminación de participantes del estudio una vez analizados los resultados, excesivas variables dependientes y eliminación de resultados negativos. Por otra parte, Blanco, Perales y Vadillo **(2017)** consideran que la falta de conocimientos estadísticos y los sesgos individuales en la toma de decisiones forman parte de la abundancia de falsos-positivos en la literatura psicológica, contribuyendo a esta abundancia las presiones editoriales y académicas en publicar resultados centrándose principalmente en el p-valor.

Tres propuestas para la evaluación cuantitativa de programas de intervención

La Open Science Collaboration **(2015)** realizaron 100 réplicas a estudios en psicología publicados, escogidos aleatoriamente, contando con la colaboración de los autores originales en el procedimiento. De los 100 estudios originales, 97 reportaban resultados significativos. Sin embargo, sólo 45 replicaciones (39%) lograron obtener este tipo de resultados; además que, de esos 45 estudios replicados, sólo 21 obtuvieron un tamaño del efecto dentro de los intervalos de confianza que reportaban los estudios originales. En otras palabras, existe un serio sesgo de publicación en las revistas de psicología ya que Fanelli **(2010)** reportaba que el 91.5% de las replicaciones publicadas eran exitosas, pero después la Open Science nos muestra que en esa cifra puede haber cerca de 52% de falsos-positivos. Lo más preocupante es que de esos trabajos publicados es donde se obtiene la información para elaborar las guías de práctica profesional (Psicología basada en evidencia), y los alcances de los meta-análisis no alcanzan a dar cuenta de los falsos-positivos, ya que se pueden considerar sólo como una herramienta para sintetizar los resultados de un conjunto de trabajos de una misma línea de investigación. En este sentido, en este apartado se presentarán tres propuestas para la evaluación cuantitativa de programas de intervención, volviendo al énfasis respecto que la evaluación de programas debería ser un paso esencial antes de emprender un trabajo de replicación.

¿Qué es la probabilidad?

Podemos entender a la teoría de la probabilidad como el estudio de eventos aleatorio con base en postulados matemáticos **(Athreya, 2015)**, considerando dos categorías: Fenómenos no aleatorios (deterministas) y aleatorios (estocásticos). Esta última sería la que nos ocuparía en psicología. A manera muy resumida, el autor nos presenta

el modelo de Kolmogorov para el estudio de fenómenos aleatorios donde el propósito es determinar, con base a un espacio muestral, los posibles resultados en función de una probabilidad dada.

A pesar de que se habla de una teoría de la probabilidad, desde la filosofía se han abordado diferentes tipos de probabilidades considerando en qué consisten **(Good, 1959).** Este autor, a pesar de describir cinco tipos de probabilidad (véase la tabla 1), argumenta que la probabilidad subjetiva debería ser suficiente. También estaríamos de acuerdo con esto ya que este tipo de probabilidad permitiría un buen acoplamiento epistemológico con el realismo crítico (descrito brevemente en apartados anteriores), y con la acumulación de nuevos conocimientos, aspecto saludable para cualquier disciplina científica.

Tabla 1. Tipos de probabilidad

Tipo de probabilidad	Descripción
Probabilidad clásica	La proporción de ocurrencia de casos igualmente probables.
Probabilidad subjetiva	Probabilidad que puede aumentar, o disminuir, dependiendo de la información que se tenga de un evento específico.
Probabilidad física	Probabilidad de realizar un evento "exitoso", dado su diseño. Pueden ser probabilidad "Verdadera" o "Hipotética"
Probabilidad inversa	Probabilidad expresada en términos de "probabilidad inicial", "probabilidad final" y verosimilitud probabilística. Los principios son parecidos a la probabilidad subjetiva.
Probabilidad frecuentista	Probabilidad calculada con base a una secuencia de eventos realizados.

Fuente: Good (1959)

Una de las aplicaciones que se le ha dado a la probabilidad es el estudio de eventos de cola **(Tail events, Barberis, 2013).** Estos son considerados eventos de alto impacto, pero de ocurrencia poco frecuente, centrados para su estudio en los pesos de probabilidad. Básicamente, se centra en valoración probabilística de la ocurrencia de un evento para posteriormente tomar una decisión al respecto.

Con lo anterior, empezamos a comprender que la probabilidad nos brinda herramientas para solventar problemáticas a través de la consideración de eventos condicionales y grados de creencias. El trabajo sobre probabilidad lógica mental **(Pfeifer, 2013)**, nos brinda la oportunidad de un primer acercamiento a la propuesta bayesiana, a pesar de que éste no es central en la propuesta de Pfeifer, que será discutida en la siguiente subsección. Él estudia la coherencia de la transición entre las premisas inciertas hacia las conclusiones (p. 330). Las probabilidades condicionadas **(Gilio y Over, 2012)** las podemos entender como la probabilidad de que un evento A ocurra dado un evento B, simbolizado P (A | B). Teniendo en cuenta esto, se podría considerar que ya se conocen los postulados básicos para proseguir con las propuestas para la evaluación de programas de intervención que serán presentadas a continuación.

La propuesta Bayesiana

Antes de comenzar a trabajar con la teoría de Bayes, deberíamos ser capaces de reconocer que la ocurrencia de los eventos a estudiar está relacionada de alguna forma. En este sentido, proponemos la estadística frecuentista como un primer paso para llevar a cabo la teoría de Bayes. Recordemos que la primera nos informa acerca de la frecuencia de ocurrencia, mientras que la segunda nos informa sobre la probabilidad de ocurrencia. En este sentido, la regla de Bayes **(Berry, 1995)** nos indica el cambio en la probabilidad de ocurrencia a medida que se obtiene mayor información sobre un evento **(p. 147).** Para una descripción metodológica más exhaustiva véase Berry **(1995, pp. 124-164).**

Correa y Barrera **(2018)** nos presentan el teorema de Bayes con la fórmula (1).

$$P(B_i|A) = (P(B_i \cap A))/(P(A)) = (P(A|B_i)P(B_i))/(\Sigma^{k}_{i=1} P(A|B_i)P(B_i)), si\ P(A) \neq 0, P(B_i) \neq 0, i = 1,2,\ldots,k \quad (1)$$

De igual forma, el cálculo para variables aleatorias según el teorema de Bayes está en (2).

$$\xi(\theta|x) = (\int(x|\theta)\,\xi(\theta))/(\int \theta f\,(x|\theta)\,\xi(\theta)\,d\theta) \quad (2)$$

En donde:

x: datos.

θ: parámetro desconocido.

f: verosimilitud de los datos dado el parámetro desconocido.

ξ(θ): distribución a priori de *θ*.

En forma resumida, el teorema de Bayes puede realizarse mediante la fórmula (3)

$$P(A|B) = (P(B|A) \cdot P(A)) / (P(B)) \quad (3)$$

Donde

P(B): Datos observados

P(B|A): Probabilidad de veracidad de los datos producidos por el evento observado

P(A): Probabilidad a priori

P(A|B):Probabilidad a posteriori.

RIP 14 (1) pág. 93-104
Revista Iberoamericana de Psicología
ISSN-L: 2027-1786

Una vez conseguido varios eventos (replicaciones), podemos realizar el mismo cálculo para ambos, por ejemplo *P(A1|B)* para el trabajo original y *P(A2|B)* para la primera réplica.

Ahora podremos comparar el factor de Bayes considerando dos probabilidades a posteriori con la siguiente fórmula (4)

$$(P(A_1|B))/(P(A_2|B)) = (P(B|A_1))/(P(B|A_2)) \times (P(A_1))/(P(A_2)) \quad (4)$$

De la fórmula anterior, la división izquierda son las probabilidades posteriores (qué tan probable es rechazar H0), la división del extremo derecho son las probabilidades a priori (las cuales pueden ser obtenidas por estudios anteriores sobre el objeto de estudio, o por alguna inferencia justificada del investigador), y la división central es el factor Bayes combinado de ambos eventos (replicaciones), por lo que nos queda la fórmula (5)

$$BF_{12} = (P(B|A_1))/(P(B|A_2)) \quad (5)$$

A su vez, Jeffreys **(1998),** presenta una tabla de criterios de decisión respecto al apoyo que ofrece el factor de Bayes en la siguiente tabla:

Tabla 2

Factor Bayes (K)	Decisión
0	Hipótesis nula se sostiene
0 a 5	Evidencia contra H0, pero apenas para mencionar
>5 a 10	Evidencia sustantiva contra H0
>10 a 15	Evidencia fuerte contra H0
>15 a 20	Evidencia muy fuerte contra H0
>20	Evidencia decisiva contra H0

Fuente: Modificación y traducido de Jeffreys (1998)

Para una instrucción más detallada, así como el manejo de Rstudio para llevar a cabo estadística bayesiana, véase Correa y Barrera **(2018).**

En términos de una psicología aplicada del teorema de Bayes, podemos replantear un ejemplo brindado por Canals **(2019).** Supóngase que la partición del espacio muestral, *P(A),* sea un diagnóstico planteado (por ejemplo, burnout presente en una población), mientras que el evento P(B) sea la presencia de síntomas asociados a este fenómeno (Despersonalización, cansancio emocional o falta de logro profesional). Entonces, la probabilidad condicional (el grado de evidencia de la presencia de burnout en una población dado que muestran esos síntomas) puede resolverse mediante la ecuación (1), permitiéndonos conocer la verosimilitud de nuestra hipótesis respaldada por la evidencia. Este planteamiento nos ofrece un parámetro para considerar que, si bien es cierto que tenemos grados de certeza respecto a las evaluaciones que realizamos, no dejamos de considerar que las personas razonan y actúan bajo condiciones de incertidumbre **(Taborda, 2009),** por lo que se estaría rechazando una concepción unívoca y determinista de la conducta humana. Esto último gracias a que, a diferencia de la estadística frecuentista, la propuesta Bayesiana no busca rechazar o aceptar la H0, sino encontrar el grado de apoyo para tomar una decisión con base a la información observada ya que es un método comparativo **(Díaz y Batanero, 2006).**

Tamaño del efecto

Otra propuesta para considerar en la evaluación de programas de intervención y replicación, es el tamaño del efecto. Al momento de hablar sobre la replicabilidad, hacíamos referencia a este parámetro, el cual es definido, según Sullivan y Feinn **(2012),** como la magnitud de la diferencia entre grupos. Según estos autores, la relevancia de este parámetro radica en que, si bien el p-valor nos informa de la existencia de un efecto, éste no nos informa su magnitud. En otras palabras, la literatura psicológica, como lo veíamos en el trabajo de la Open Science Colaboration **(2015),** páginas arriba, está plagada de trabajos con resultados estadísticamente significativos, sin embargo, estimar el tamaño del efecto es útil al momento de considerar la importancia práctica y teórica de un trabajo **(Fritz, Morris y Richler, 2012**), permitiendo una mejor interpretación y descripción de los resultados. Estos autores mencionaban que la presencia de efectos grandes, pero no significativos, implicaría poca potencia estadística; mientras que efectos pequeños, pero significativos, serían una advertencia de un posible error tipo 1. Una vez descrito la importancia de este parámetro, pasemos a la descripción de su cálculo.

Primero hay que remarcar que el tamaño del efecto está dividido en dos grupos **(Rosenthal, 1994, como se citó en Lakens, 2013)**: la familia d (descripción de las diferencias entre observaciones) y la familia r (descripción de la proporción de la varianza explicada por miembros de un grupo).

El parámetro d de Cohen muestra las diferencias de medias estandarizadas entre dos grupos, calculada mediante la formulada (6)

$$ds = (\bar{X}1 - \bar{X}\,2)/\sqrt{((n1 -1)SD^2_1+(n2 -1)SD^2_2)/(n1 +n2 -2))} \quad (6)$$

donde:

X: Media observada.

SD^2: Desviación estándar al cuadrado por grupo.

n: número de observaciones por grupo.

La interpretación del coeficiente de Cohen se da en la siguiente tabla:

Tabla 3

Coeficiente	Interpretación
0 – 0.2	Efecto pequeño
0.21 – 0.8	Efecto mediano
>0.8	Efecto grande

Fuente: Lakens (2013)

Por su parte, para describir utilizaremos la prueba de eta cuadrada parcial, siendo ésta una prueba expandida de dicho parámetro. Esta prueba mide la proporción de la varianza de la variable dependiente asociada a una variable independiente, siendo “la suma de los cuadrados del efecto dividido por el número total de cuadrados” **(Lakens, 2013, p. 5).** La fórmula se muestra en (7)

$$\eta^2 = SS_{effect}/SS_{total} \quad (7)$$

RIP 14 (1) pág. 93-104
Revista Iberoamericana de Psicología
ISSN-L: 2027-1786

donde:

SSeffect: Suma de los cuadrados.

SStotal: Suma de los cuadrados totales.

El resultado de esta división siempre nos dará *<1,* puesto que el decimal resultante se interpretaría tal que *η2 = .13* representa que el 13% de la varianza puede ser atribuida al grupo, o condición, al que se pertenece.

Poder calcular el tamaño del efecto en los programas de intervención, orientaría la toma de decisión **(Rhea, 2004)** sobre la efectividad que hayan tenido en una población, puesto que nos informa acerca de qué tanta repercusión podríamos esperar de llevarlo a cabo.

En psicología, las evaluaciones sobre los resultados reportados mediante este parámetro son realizado en importantes revistas **(Castillo-Blanco y Alegre, 2015)** para ajustar el análisis de resultados a la práctica de los autores; sin embargo, se reporta una escaza utilización de este parámetro en las publicaciones latinoamericanas. En un contexto más general, las consideraciones respecto a la integración del tamaño del efecto en las publicaciones en psicología en español comenzaron a realizarse a partir de 1995 **(García, Ortega y De la Fuente, 2011)**. Estos autores revisaron las publicaciones de cuatro revistas durante un periodo de seis años, donde el reporte del tamaño del efecto no rebasó el **31.8%** de publicaciones, incluso llegando a **2.9%** en una revista, esto a pesar de que la Asociación Americana de Psicología, desde el 2001, considera un defecto en el diseño y reporte de investigación el no incluirlo. Por lo que se considera pertinente la difusión de este tipo de análisis de los resultados en el contexto.

Intervalos de confianza

De igual manera, la APA **(2001),** consideraba como un defecto en el análisis estadístico no incluir los intervalos de confianza (IC). Utilizar intervalos de confianza nos permitiría conocer la diferencia del efecto del programa, así como determinar la precisión de la estimación de dicha diferencia **(Young y Lewis, 1997a).** En este sentido, los intervalos de confianza son un rango de posibles valores resultantes de un evento, asignados con base en un margen de error. Así, la longitud de los IC nos permite inferir la precisión de los resultados, donde mientras menor sea el intervalo entre los rangos, mayor será la precisión, y viceversa. La cuestión con los IC es que los rangos obtenidos son consistentes con las observaciones realizadas, por lo que se esperaría que en replicaciones se obtuvieran, no resultados idénticos al original, sino resultados que entren dentro del rango establecido en el estudio original.

Comúnmente, los IC se calculan a través de una distribución t, para dos situaciones: Considerando una media muestral (Intervalos grupal) y considerando dos medias muestrales (Intervalos inter-grupales). Sin embargo, estas no son los únicos cálculos con IC **(véase Young y Lewis, 1997b).**

Para calcular los IC de una sola media muestral se utilizaría la fórmula (8)

$$\bar{X} \pm (t2\alpha \cdot S/\sqrt{n}) \quad (8)$$

donde:

X: Media de las observaciones.

S: Desviación estándar.

√n: Raíz cuadrada del tamaño de la muestra.

t2α: Valor de una cola de t-Student con distribución *n-1* df.

Recuérdese que el símbolo *"+"* significa que esta operación debe realizarse primero en sumatoria para obtener el valor máximo del intervalo, y luego como resta para obtener el valor mínimo del intervalo.

Por su parte, los IC considerando dos medias muestrales se obtienen mediante la fórmula (9)

$$(\bar{x}1) - (\bar{x}2) \pm (t2\alpha: \cdot S\sqrt{((1/n1) + (1/n2))} \quad (9)$$

donde:

x1: Media del grupo 1.

x2: Media del grupo 2.

t2α:: Valor de una cola de t-Student con distribución *n-1* df.

S: Desviación estándar.

n1: Número de observaciones del grupo 1.

n2: Número de observaciones del grupo 2.

Una aplicación que se le atribuye a los IC es la capacidad de ser considerados como Intervalos Predictivos. La interpretación de los IC brinda la base para pensarlos en varios niveles **(Cumming y Fidler, 2018)** donde futuras replicaciones de un estudio obtengan un resultado entre el rango de los IC del estudio original. La ventaja de esta aplicación es parecida a lo expuesto en el teorema de Bayes, ya que los IC pueden acumularse para dar cuenta de un IC actualizable en relación con la media *(μ).*

Otra aplicación de los parámetros de IC la brindan Domínguez-Lara y Merino-Soto **(2015)** para el área de psicometría argumentan su uso para conocer la variación del coeficiente de Cronbach esperado dentro de un parámetro poblacional, aspecto de relevancia al momento de realizar diagnósticos clínicos o comunitarios. Por su parte, Cumming y Finch **(2005)** argumentan cuatro ventajas del uso de los IC, de los cuales mencionamos tres: (1) brindan una mejor comprensión de la situación del estudio, (2) su interpretación es sencilla si se tiene experiencia con los p-valores, (3) ayudan a la acumulación del conocimiento de una disciplina.

Discusión

Evaluación y replicabilidad: Cuestiones a considerar

Mediante un repaso acerca de qué es la evaluación, y el proceso para su realización, se ha considerado plantear que no se puede hablar de una "crisis de la replicabilidad", sin tomar en cuenta procesos de evaluación del trabajo original como error muestral, validez, efecto observado e incluso aspectos contextuales **(Stanley y Spence, 2014).** Este planteamiento estaría acorde con autores **(Tressoldi, 2012; Etz y Vandekerckhove, 2016)** quienes cuestionan si esta problemática es debido al evasivo objeto de estudio de la psicología o a una falta de medidas implementadas para aumentar la potencia estadística; disminuida por el abuso, y malinterpretación, de la estadística

RIP 14 (1) pág. 93-104
Revista Iberoamericana de Psicología
ISSN-L: 2027-1786

frecuentista. Para lidiar con esta falta de potencia estadística, se han realizado tres propuestas que, de acuerdo con lo revisado en los apartados de "Evaluación de programas de intervención" y "Crisis de la replicabilidad en psicología", podría beneficiar la práctica profesional al fomentar no sólo la acumulación del conocimiento, sino a su actualización continua. Por ello, se quiere hacer una propuesta aún más general, que es adoptar la estadística Bayesiana como un nuevo paradigma en la investigación en psicología, o por lo menos como paradigma complementario al actual, puesto que este tipo de procedimiento englobaría conocer el grado de certeza que plantean el tamaño del efecto e intervalos de confianza de parámetros tradicionalmente utilizados, tanto para el estudio original como para la replicación **(Verhagen y Wagenmakers, 2014).** Así pudiéramos solventar las deficiencias que las pruebas de significación de hipótesis nulas centradas en el p-valor han evocado.

Por último, un aspecto a considerar en este trabajo es que, si bien el autor se ha centrado en parámetros cuantitativos considerados de utilidad para la evaluación de programas de intervención a raíz de lo que se esperaría de una práctica científica, no se está declarando que dicha aproximación sea la única, o la mejor herramienta, que poseemos para solventar la crisis de la replicabilidad. Se conoce la enorme contribución que, desde los métodos cualitativos **(Hendrick, 1990; Park, 2004)** se ha brindado a la replicabilidad y se esperaría que en un futuro ambos métodos puedan considerarse de forma conjunta para el diseño de replicaciones y el análisis de resultados. Sin embargo, debido a las limitaciones de extensión de este trabajo, se ha decidido enfocarse en métodos cuantitativos, esperando que en futuros trabajos se logren avances en la formalización de enfoques.

Con este trabajo de reflexión se buscó vislumbrar ciertos aspectos respecto al trabajo de evaluación y replicación de programas de intervención en psicología tratando de establecer de dónde surgen ambas actividades; además, se trató de esclarecer el rol que han tenido en la llamada "crisis de replicabilidad" en psicología. Para hacer frente a esta crisis, no solamente nos tenemos que centrar en los aspectos metodológicos que plantean los trabajos de replicación, sino, también, aquellos aspectos que plantea la evaluación de los programas de intervención, puesto que, como se mencionó, una intención de abordar ambos temas en el mismo trabajo fue que se considera al último un criterio necesario para establecer la viabilidad de llevar a cabo al primero.

Sin duda, cabría esperar que este nuevo evento dentro del trabajo profesional de la psicología no sólo es una cuestión de sofisticación metodológica, sino también teórica, por lo que es una oportunidad para realizar una reflexión crítica acerca de qué estamos haciendo y cómo lo estamos haciendo. Esto último debido a que sería ingenuo creer que se obtendrían mejores resultados mientras mayor número de parámetros estadísticos se contemplen en la labor del psicólogo. Esto representaría poca ayuda si no existe un refinamiento teórico que sustente su aplicación y una valoración positiva hacia estos para que estudiantes de pregrado se interesen en la replicación de estudios y se involucren en el mundo científico **(Cota, Beltrán, Tánori y Vázquez, 2019).**

Referencias

American Psychological Association. (2001). Publication manual of the American Psychological Association (5th ed.). APA.

American Psychological Association. (2006). Evidence-Based Practice in Psychology. American Psychologist, 61(4), 271-285.Doi: *10.1037 / 0003-066X.61.4.271*

Athreya, K. B. (2015). What is probability theory? Resonance, 20(4), 292-310. Doi: *https://doi.org/10.1007/s12045-015-0186-3*

Barberis, N. (2013). The psychology of tail events: Progress and challenges. American Economic Review: Papers & Proceedings, 103(3), 611-616. *https://www.aeaweb.org/articles?id=10.1257/aer.103.3.611*

Berry, D. (1995). Statistics: A bayesian perspective. Duxbury Press. *https://www.jstor.org/stable/2684909?seq=1*

Blanco, F., Perales, J. C. y Vadillo, M. A. (2017). Pot la psicologia rescatar-se a si mateixa? Incentius, biaix i replicabilitat. Anuari de psicologia de la Societat Valenciana de Psicologia, 18(2), 231-252. *https://digibug.ugr.es/bitstream/handle/10481/49803/Blanco_Replicabilidad.pdf?sequence=1*

Bollen., K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A. y Old, J. L. (2015). Social, behavioral, and economic sciences perspectives on robust and reliable science. National Science Fundation. Recuperado de: https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf.

Brandt, M. J., Ijzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R. y Veer, A. (2014). The replication recipe: What makes for a convincing replication? Journal of Experimental Social Psychology, (50), 217-224. *https://psycnet.apa.org/record/2013-13974-014*

Bunge, M. (1983). Treatise on basic philosophy Volumen 5 Epistemology and methodology I: Exploring the world. Netherlands: D. REIDEL PUBLISHING COMPANY.

Bunge, M. (1993). Realism and antirealism in social science. Theory and Decision, 35, 207-235. *https://link.springer.com/article/10.1007/BF01075199*

Bunge, M. (2012). Tratado de Filosofía Volumen 4 Ontología II: Un mundo de sistemas. España, Gedisa.

Canals, M. (2019). Bases científicas del razonamiento clínico: inferencia Bayesiana. Revista Médica de Chile, 147, 231-237. *https://scielo.conicyt.cl/scielo.php?pid=S0034-98872019000200231&script=sci_arttext*

Castillo-Blanco, R. W. y Alegre, A. A. (2015). Importancia del tamaño del efecto en el análisis de datos de investigación en psicología. Persona, (18), 137-148. *https://dialnet.unirioja.es/servlet/articulo?codigo=6112633*

Cohen, J. (1994). The earth is round (p <.05). American Psychological Association, 49(12), 997-1003. *https://psycnet.apa.org/record/1995-12080-001*

Cook, J. R. (2014). Using evaluation to effect social change: Looking through a community psychology lens. American Journal of Evaluation, 36(1), 107-117. *https://journals.sagepub.com/doi/10.1177/1098214014558504*

Correa, J. C. y Barrera, C. J. (2018). Introducción a la Estadística Bayesiana. Medellín, Colombia: Fondo Editorial ITM. *https://repositorio.itm.edu.co/handle/20.500.12622/1793*

Cota, L., Beltrán, J., Tánori, J. y Vázquez, M. (2019). Propiedades psicométricas de una escala de actitudes hacia la investigación científica (EACIN): Estudio en alumnos universitarios mexicanos. Revista Iberoamericana de Psicología, 12(3), 43-54. *https://www.researchgate.net/profile/Jesus-Beltran-Sanchez/publication/341722287_Propiedades_psicometricas_de_una_escala_de_actitudes_hacia_la_investigacion_cientifica_EACIN_Estudio_en_alumnos_universitarios_mexicanos/links/5fa5a042a6fdcc06241cb713/Propiedades-psicometricas-de-una-escala-de-actitudes-hacia-la-investigacion-cientifica-EACIN-Estudio-en-alumnos-universitarios-mexicanos.pdf*

Crandall, C. S. y Sherman, J. W. (2015). On the scientific superiority of conceptual replications for scientific progress. Journal of Experimental Social Psychology, 66, 93-99. *https://www.sciencedirect.com/science/article/abs/pii/S0022103115300020*

Cumming, G. y Fidler, F. (2018). Confidence intervals: Better answers to better questions. Zeitschrift für Psychologie / Journal of Psychology, 217(1), 15-26. *https://www.nature.com/articles/s41598-020-78438-4*

Cumming, G. y Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. American Psychologist, 60(2), 170-180. https://psycnet.apa.org/record/2005-01817-003

Dahiru, T. (2008). P-value, a true test of statistical significance? A cautionary note. Annals of Ibadan Postgraduate Medicine, 6(1), 21-26. https://www.ajol.info/index.php/aipm/article/view/64038

Díaz, C. y Batanero, C. (2006). ¿Cómo puede el método Bayesiano contribuir a la investigación en psicología y educación? Paradigma, 27(2), 35-54. http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1011-22512006000200003

Domínguez-Lara, S. A. y Merino-Soto, C. (2015). ¿Por qué es importante reportar los intervalos de confianza del coeficiente alfa de Cronbach? Revista Latinoamericana de Ciencias Sociales, Niños y Juventud, 13(2), 1326-1328. https://www.redalyc.org/pdf/773/77340728053.pdf

Donaldson, S. I., Gooler, L. E. y Scriven, M. (2002). Strategies for managing evaluation anxiety: Toward a psychology of program evaluation. American Journal of Evaluation, 23(3), 261-273. https://journals.sagepub.com/doi/abs/10.1177/109821400202300303

Earp, B. D. y Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. Frotiers in Psychology, 6. 1-11. Doi: 10.3389/fpsyg.2015.00621. https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00621/full

Echeburúa, E., Corral, P. y Salaberria, K. (2010). Efectividad de las terapias psicológicas. Un análisis de la realidad actual. Revista de Psicopatología y Psicología Clínica, 15(2), 85-99. http://e-spacio.uned.es/fez/eserv/bibliuned:Psicopat-2010-15-2-5010/Documento.pdf

Etz, A. y Vandekerckhove, J. (2016). A bayesian perspective on the reproductibility Project: Psychology. Plos one, 11(2), 1-12. doi: 10.1371/journal.pone.0149794. https://journals.lww.com/pedpt/Fulltext/2014/26010/Motor_Competence_and_Physical_Fitness_in.13.aspx

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. Plos One, 5(4). 1-10. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0010068

Fernández-Ballesteros, R. (2013a). Conceptos y modelos básicos. En R. Fernández-Ballesteros. (Dir.), Evaluación psicológica: conceptos, métodos y estudio de casos. Madrid, España: Pirámide. http://www.kydconsultores.com/shared_books/001-EP-RFB.pdf

Fernández-Ballesteros, R. (2013b). El proceso como procedimiento científico y sus variantes. En R. Fernández-Ballesteros. (Dir.), Evaluación psicológica: conceptos, métodos y estudio de casos (pp. 27-59). Madrid, España: Pirámide. https://dialnet.unirioja.es/servlet/articulo?codigo=4924575

Francis, G. (2012). The psychology of replication and replication in psychology. Perspective on Psychological Sciences. 7(6), 585-594. doi: 10.1177/1745691612459520.

Fritz, C. O., Morris, P. E. y Richler, J. J. (2012). Effect size estimates: Current use, calculation, and interpretation. Journal of Experimental Psychology: General, 141(1), 2-18. doi: 10.1037/a0024338.

Gairín, J. (2010). La evaluación del impacto en programas de formación. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 8(5), 19-43. https://revistas.uam.es/reice/article/view/4724/5158

García, J., Ortega, E. y De la Fuente, L. (2011). The use of the effect size in JCR spanish journals of psychology: From theory to fact. The Spanish Journal of Psychology, 14(2), 1050-1055. doi: 10.5209/rev_SJOP.2011.v14.n2.49.

García-Garzón, E., Lecuona, O. y Carbajal, G. V. (2018). Estudios de replicación, pre-registros y ciencia abierta en psicología. Apuntes de Psicología, 36(1-2), 75-83.

Gilio, A. y Over, D. (2012). The psychology of inferring conditionals from disjunctions: A probabilistic study. Journal of Mathematical Psychology, 56, 118-131. doi: 10.1016/j.jmp.2012.02.006.

Good, I. J. (1959). Kinds of probability. Science, 129(3347), 443-447. https://www.jstor.org/stable/1757847

Goodman, S. N., Fanelli, D. y Ioannidis, J. P. (2016). What does research reproducibility mean? Science Translational Medicine, 8(341), 1-6. https://stm.sciencemag.org/content/8/341/341ps12.short

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: Are they important? Journal of Social Behavior & Personality, 5(4), 41–49. https://search.proquest.com/openview/1561d1420ebe8c77538bce783314dac5/1?pq-origsite=gscholar&cbl=1819046

Hubbard, R. y Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. Theory & Psychology, 18(1), 69-88. https://doi.org/10.1177/0959354307086923

Hunsley, J. (2007). Training psychologists for evidence-based practice. Canadian Psychology/Psychologie Canadienne, 48(1), 32–42. doi: 10.1037/cp2007005.

Jeffreys, H. (1998). The theory of probability. United Kingdom: Oup Oxford.

Kazdin, A. E. (1986). Comparative outcome studies of psychotherapy: Methodological issues and strategies. Journal of Consulting and Clinical Psychology, 54(1), 95–105. https://doi.org/10.1037/0022-006X.54.1.95

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. American Psychologist, 63(3), 146–159. doi: 10.1037/0003-066X.63.3.146.

Kelly, J. G. (1966). Ecological constraints on mental health services. American Psychological Association, 21(6), 535-539. https://doi.org/10.1037/h0023598

Koole, S. L. y Lakens, D. (2012). Rewarding replications: a sure and simple way to improve psychological science. Perspective on Psychological Science, 7(6), 608-614. doi: 10.1177/1745691612462586

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-test and ANOVAs. Frontiers in Psychology, 4(863), 1-12. doi: 10.3389/fpsyg.2013.00863.

Ludwig, J. L., Kling, J. R. y Mullainathan, S. (2011). Mechanism experiments and policy evaluations. Journal of economic Perspectives, 25(3), 17-38. doi: 10.1257/jep.25.3.17.

Makel, M. C y Plucker, J. A. (2014a). Creativity is more than novelity: Reconsidering replication as a creativity act. Psychology of Aesthetic, Creativity and the Arts, 8(1), 27-29. doi: 10.1037/a0035811.

Makel, M. C. y Plucker, J. A. (2014b). Facts are more important than novelty: Replication in the education sciences. Educational Researcher, 20(10), 1-13. doi: 10.3102/0013189X14545513.

Makel, M. C., Plucker, J. A. y Hegarty, B. (2012). Replications in psychology research: How often do they really occur? Perspectives on Psychological Science, 7(6), 537-542. doi: 10.1177/1745691612460688.

Martorell, C. y Gómez, O. (2010). Enfoque de la evaluación psicológica de la revista iberoamericana de diagnóstico y evaluación psicológica (Ridep). Revista Iberoamericana de Diagnóstico y Evaluación, 2(30), 35-55. https://www.redalyc.org/pdf/4596/459645442003.pdf

Maxwell, S. E., Lau, M. Y. y Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? American Psychological Association, 70(6), 487-498. doi: 10.1037/a0039400.

Miller, R. L. (2014). Community psychology, evaluation, and social critique. American Journal of Evaluation, 36(1), 89-99. doi: 10.1177/1098214014557694.

Moriana, J. A. y Martínez, V. A. (2011). La psicología basada en evidencia y el diseño y evaluación de tratamientos psicológicos eficaces. Revista de Psicopatología Clínica, 16(2), 81-100. http://revistas.uned.es/index.php/RPPC/article/view/10353

National Center for Chronic Disease Prevention and Health Promotion. (2011). Developing an effective evaluation plan. Setting the course for effective program evaluation. Recuperado de: https://www.cdc.gov/obesity/downloads/cdc-evaluation-workbook-508.pdf.

Newman, D. L., Scheirer, M. A., Shadish, W. R. y Wye, C. (1995). Guiding principles for evaluators. New directions for program evaluation, (66), 19-26. https://doi.org/10.1002/ev.1706

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Nature, 349(6251), aac4716–aac4716. doi: 10.1126/science.aac4716.

Ordoñez, O. (2014). Replicar para comprender: prácticas investigativas para promover el razonamiento científico en estudiantes de psicología. Pensamiento Psicológico, 12(2), 7-24. doi: 10.11144/Javerianacali. PPSI12-2.rcpi.

Ortegón, E., Pacheco, J. F. y Prieto, A. (2015). Metodología del marco lógico para la planificación, el seguimiento y la evaluación de proyectos y programas. Santiago de Chile, Chile: Naciones Unidas. http://148.202.167.116:8080/xmlui/bitstream/handle/123456789/3839/Metodolog%c3%ada_del_marco_l%c3%b3gico.pdf?sequence=1&isAllowed=y

Park, C. L. (2004). What is the value of replicating other studies? Research Evaluation, 13(1), 189-195. https://academic.oup.com/rev/article-abstract/13/3/189/1538811

Pfeifer, N. (2013). The new psychology of reasoning: A mental probability logic perspective. Thinking & Reasoning, 19(3-4), 329-345. doi: 10.1080/13546783.2013.838189.

Rhea, M. R. (2004). Determine the magnitude of treatment effects in strength training research through the use of effect size. Journal of Strength and Conditioning Research, 18(4), 918-920. file:///C:/Users/09/Downloads/2004-Rhea-EFFECTSIZEFuerza.pdf

Rossi, P. H. (1987). The iron law of evaluation and other metallic rules. Research in Social Problems and Public Policy, 4, 3-20. https://www.gwern.net/docs/sociology/1987-rossi

Santos, D. (2013). La crisis en la psicología social contemporánea: el fenómeno del priming. Revista Electrónica de Psicología Social «Poiésis», 25, 1-8. https://www.funlam.edu.co/revistas/index.php/poiesis/article/view/636

Schmidt, F. L. y Oh, I. S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? Archives of Scientific Psychology, 4, 32-37. doi: 10.1037/arc0000029

Sheldon, J. A. y Wolfe, S. M. (2014). The community psychology evaluation nexus. American Journal of Evaluation, 36(1), 86-117. doi: 10.1177/1098214014558503.

Shipman, S. (1989). General Criteria for evaluating social programs. Evaluation Practice, 10(1), 20–26. https://doi.org/10.1177/109821408901000104.

Simons, D. J. (2014). The value of direct replication. Perspectives on Psychological Science, 9(1), 76-80. doi: 10.1177/1745691613514755.

Singh, K., Ang, S. H. y Leong, S. M. (2003). Increasing replication for knowledge accumulation in strategy research. Journal of Management, 29(4), 533-549. doi: 10.1016/S0149-2063(03)00024-2.

Spence, J. R. y Stanley, D. J. (2016). Prediction interval: What to expect when you're expecting… A replication. Plos One. 11(9), 1-22. doi: 10.1371/journal.pone.0162874.

Stanley, D. J. y Spence, J. R. (2014). Expectations for replications: Are yours realistic? Perspectives on Psychological Science, 9(3), 305-318. doi: 10.1177/1745691614528518.

Sullivan, G. M. y Feinn, R. (2012). Using effect size-or why p value is not enough. Journal of Graduate Medical Education, 4(3), 279-282, doi: 10.4300/JGME-D-12-00156.1.

Taborda, H. (2009). Modelos bayesianos de inferencia psicológica: ¿cómo predecir acciones en situaciones de incertidumbre? Universitas Psychologica, 9(2), 495-507. file:///C:/Users/09/Downloads/420-Texto%20del%20art%C3%ADculo-2413-1-10-20100616.pdf

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F. y Shrout, P. E. (2017). It's time to broaden the replicability conversation: thoughts for and from clinical psychological science. Perspectives on Psychoogical Science, 12(5), 742-756. doi: 1oi.org/10.1177/1745691617690042

Tressoldi, P. E. (2012). Replication unreliability in psychology: elusive phenomena or "elusive" statistical power? Frontiers in Psychology, 3(218), 1-5. doi: 10.3389/fpsyg.2012.00218.

Tsang, E. W. y Kwan, K. M. (1999). Replication and theory development in organizational science: a critical realist perspective. Academy of Management Review, 24(4), 759-780. https://doi.org/10.5465/amr.1999.2553252

Verhagen, J. y Wagenmakers, E. J. (2014). Bayesian test to quantify the result of a replication attempt. Journal of Experimental Psychology: General, 143(4), 1457-1475. doi: 10.1037/a0036731.

World Health Organization. (2013). WHO evaluation practice handbook. Recuperado de: https://apps.who.int/iris/bitstream/handle/10665/96311/9789241548687_eng.pdf;jsessionid=5F26E0816B02FB4EEF5FB4D4D4C4E484?sequence=1

Ybarra, J., Orozco, L. y Valencia, A. (2015). Tratamientos psicológicos con apoyo empírico y práctica clínica basada en la evidencia. En J. Ybarra, L. Orozco y A. https://www.researchgate.net/publication/295851247_Tratamientos_psicologicos_con_apoyo_empirico_y_practica_clinica_basada_en_la_evidencia

Valencia. (Coords.), Intervenciones con apoyo empírico: herramienta fundamental para el psicológo clínico y de la salud (pp. 1-30). México: Manual Moderno.

Young, K. D. y Lewis, R. J. (1997a). What is confidence? Part 1: The use and interpretation of confidence intervals. Annals of Emergency Medicine, 30(3), 307-310. https://doi.org/10.1016/S0196-0644(97)70166-5

Young, K. D. y Lewis, R. J. (1997b). What is confidence? Part 2: Detailed definition and determination of confidence intervals. Annals of Emergency Medicine, 30(3), 311-318. https://doi.org/10.1016/S0196-0644(97)70166-5