Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

https://doi.org/10.23913/ride.v12i23.981

Artículos científicos

Análisis de calidad de artículos educativos con diseños experimentales

Quality Analysis for Educational Articles with Experimental Designs

Análise de qualidade de artigos educacionais com projetos experimentais

Héctor Francisco Ponce Renova
Universidad Autónoma de Ciudad Juárez, México
hector.ponce@uacj.mx
https://orcid.org/0000-0002-9302-3740

Diana Irasema Cervantes Arreola
Universidad Autónoma de Ciudad Juárez, México
diana.cervantes@uacj.mx
https://orcid.org/0000-0003-2353-1309

Beatriz Anguiano Escobar
Universidad Autónoma de Ciudad Juárez, México
beatriz.anguiano@uacj.mx
https://orcid.org/0000-0002-3533-5607

Resumen

El objetivo de este trabajo fue evaluar la calidad estadística de artículos arbitrados (n = 34) con diseños experimentales y enlistar una serie de recomendaciones para incrementar su calidad. Los artículos fueron obtenidos de la Red Iberoamericana de Innovación y Conocimiento Científico (Redib) con los criterios de selección: *a)* haber llevado a cabo algún *diseño experimental* en investigación educativa y *b)* haber usado un análisis *paramétrico*. Bajo una metodología cuantitativa, exploratoria, estadística descriptiva e inferencial, se evaluó la calidad de los artículos bajo cuatro criterios: poder estadístico, susceptibilidad de réplica, acceso a bases de datos para constatar los resultados y estadísticas sugeridas por el Asociación Americana de Psicología (APA). En los resultados se encontró que la mayoría de los artículos no dieron suficiente información para apoyar sus conclusiones y el reporte de las estadísticas tuvo diferencias estadísticamente significativas y mostró un tamaño de efecto grande. Por lo tanto, se dictaminó una mala calidad. En conclusión, y para mejorar la calidad, se recomienda usar guías para los análisis estadísticos, la réplica de estudios y dar acceso a las bases de datos para que los demás autores puedan observar lo que se hizo.

Palabras clave: calidad, diseño experimental, inferencia, poder, réplica.

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

Abstract

The objective of this work was to evaluate the statistical quality of refereed articles (n = 34) with experimental designs and to list a series of recommendations to increase their quality. The articles were obtained from the Red Iberoamericana de Innovación y Conocimiento Científico (Redib) with the selection criteria: *a)* having carried out an experimental design in educational research and *b)* having used a parametric analysis. Under a quantitative, exploratory, descriptive, and inferential statistics methodology, the quality of the articles was evaluated under four criteria: statistical power, susceptibility to replication, access to databases to verify the results and statistics suggested by the American Psychological Association (APA). In the results, it was found that most of the articles did not give enough information to support their conclusions and the reporting of the statistics had statistically significant differences and showed a large effect size. Therefore, they were evaluated with poor quality. In conclusion, and to improve quality, it is recommended to use guides for statistical analysis, replication of studies and to give access to databases so that other authors can observe what was done.

Keywords: quality, experimental design, inference, power, replication.

Resumo

O objetivo deste trabalho foi avaliar a qualidade estatística de artigos referenciados (n = 34) com desenhos experimentais e listar uma série de recomendações para aumentar sua qualidade. Os artigos foram obtidos junto à Rede Ibero-Americana de Inovação e Conhecimento Científico (Redib) com os critérios de seleção: a) ter realizado um desenho experimental em pesquisa educacional eb) ter utilizado uma análise paramétrica. Sob metodologia de estatística quantitativa, exploratória, descritiva e inferencial, a qualidade dos artigos foi avaliada sob quatro critérios: poder estatístico, suscetibilidade à replicação, acesso a bancos de dados para verificação dos resultados e estatísticas sugeridas pela American Psychological Association (APA). Nos resultados, constatou-se que a maioria dos artigos não forneceu informações suficientes para embasar suas conclusões e o relato das estatísticas apresentou diferenças estatisticamente significativas e apresentou grande tamanho de efeito. Portanto, a má qualidade foi considerada. Em conclusão, e para melhorar a qualidade, recomenda-se a utilização de guias para análise estatística, replicação de estudos e acesso a bases de dados para que outros autores possam observar o que foi feito.

Palavras-chave: qualidade, desenho experimental, inferência, poder, replicação.

Fecha Recepción: Diciembre 2020 **Fecha Aceptación:** Julio 2021

Introducción

Este trabajo es una evaluación de una muestra de artículos con diseños experimentales que fueron publicados en revistas arbitradas. Haciendo una referencia a la COVID-19 y diseños experimentales, ha sido fundamental diferenciar a personas infectadas, detectar sus contactos y proveer aislamiento, así como aplicar tratamientos efectivos y vacunación. Para tener éxito, se requieren de procesos experimentales y estadísticos (entre muchos otros recursos más) para diferenciar la *efectividad* de una prueba de detección y un *efecto* (i. e., vacunas y tratamientos). Ahora bien, el método científico no solo se utiliza para solucionar *problemas de salud*, sino también para abordar problemáticas educativas (e. g., efecto de tutorías, aprendizaje a distancia, estrategias didácticas, entre otros). Al respecto, Maxwell, Delaney y Kelley (2018) explican que los "métodos de diseño experimental y análisis de datos derivan su valor de las contribuciones que hacen a la actividad en general de la ciencia" (p. 3). De una manera similar a la medicina, este texto analiza algunas partes de la calidad de los análisis estadísticos usados en procesos experimentales en investigación educativa. Feynman (1974), durante un discurso dirigido a una generación de estudiantes recién graduados, subrayó la importancia de revelar toda la información pertinente y de ser cuidadosos al hacer ciencia:

> El primer principio es que uno no debe de engañarse a sí mismo —y uno es la persona más fácil de engañar. Así que se debe de ser muy cuidadoso al respecto. Después de no haberse autoengañado, es fácil no engañar a los otros científicos. Uno solo tiene que ser honesto en la manera convencional después de todo (p. 12).

Como se verá más adelante a detalle, si un estudio no tiene el suficiente *poder estadístico* (Ponce 2019), una investigadora, al pasar por alto un efecto no detectado por la falta de poder, podría inferir que los resultados no son estadísticamente significativos y llegar a autoengañarse.

Por otro lado, el presente estudio trata de inferencias estadísticas y no de inferencias teóricas (Meehl, 1990). Las inferencias estadísticas parten de una muestra de participantes, observaciones u objetos para generalizar una serie de resultados a la población correspondiente y engloban los siguientes elementos: poder, prueba de hipótesis (e. g., test *t* o *F*), cálculo de un intervalo de confianza y estimación de un tamaño de efecto (*d* de Cohen) (Cumming y Calin-Jageman, 2017).

Se trata de ofrecer un conjunto de recursos para los y las investigadoras educativas en el área de los diseños experimentales y sus correspondientes estadísticas. La pregunta de investigación del presente estudio fue: ¿cuál ha sido la calidad del reporte de algunas estadísticas de publicaciones educativas arbitradas y relacionadas a procesos experimentales? De la pregunta se derivó uno de los objetivos: evaluar la calidad de algunos procesos estadísticos de las publicaciones educativas experimentales publicadas. El segundo objetivo fue dar recomendaciones para incrementar su calidad. Esta calidad tuvo los siguientes criterios:

1) Haber discutido, calculado y obtenido el suficiente *poder* para rechazar una hipótesis nula falsa.

2) Observar si el estudio en cuestión es parte de una serie de *réplicas*.

3) Ver si se dio *acceso* a las bases de datos para potenciales réplicas.

4) Reportar estadísticas sugeridas por la Asociación Americana de Psicología [APA, por sus siglas en inglés] (2001, 2020) desde el 2001 hasta el presente día.

Marco teórico

En este apartado se detallarán algunos aspectos relacionados con los criterios arriba mencionados. Respecto al primero de ellos, el peso del poder estadístico, la APA (2020) explica lo siguiente:

> Cuando se apliquen las estadísticas inferenciales, tomen seriamente las consideraciones relacionadas con el poder estadístico asociado con las pruebas para las hipótesis. Tales consideraciones se relacionan con la probabilidad de correctamente rechazar las hipótesis puestas a prueba dados un determinado nivel del alfa, un tamaño de efecto y un tamaño de muestra. A este respecto, hay que proveer evidencia que el estudio tiene suficiente poder para detectar efectos de interés substantivo (p. 86).

En cuanto al segundo criterio, Feynman (1974) también sugiere, en lugar de insistir en algo novedoso en cada ocasión, *repetir* experimentos para ver si se encontró la causa y el efecto de un fenómeno. Además, habría que ver si el fenómeno se repite antes de creer que se tiene algo: i. e., los valores *p* fueron conceptualizados a largo plazo y no para una sola ocasión (Greenland *et al*., 2016; Harms y Lakens, 2018).

En relación con el tercer criterio, habría que ver si se dio acceso a las bases de datos. Por ejemplo, el Center for Open Science fue creado en 2013 para propiciar la apertura, integridad y reproducibilidad de investigaciones científicas (Cumming y Calin-Jageman, 2017). Es una organización sin fines de lucro patrocinada por fondos gubernamentales y privados con aproximadamente 205 revistas suscritas. En la primera etapa, el proceso es gratuito e involucra el mandar un protocolo con una introducción, métodos y los resultados de algún pilotaje. Luego, el manuscrito es evaluado por editores y revisores, quienes pueden dar retroalimentación. Bajo la condición de seguir el protocolo en el experimento, la revista en cuestión garantiza la publicación del manuscrito. En la segunda etapa, el manuscrito deberá incluir la introducción, los métodos, los resultados de los nuevos análisis y la discusión. A los autores se les puede pedir o requerir que compartan sus *sets* de datos en un archivo público y gratuito de accesar y se les exhorta a compartir el código de sus análisis estadísticos. Después de este proceso, el artículo finalizado es publicado. Finalmente, un reporte de registro es publicado y aparecerá para dar confianza a los y las lectoras que las hipótesis y análisis principales están libres de prácticas de investigación cuestionables. Soler (2016) complementa que “dicho documento debe de contener información suficiente que permita a otros investigadores del tema entender los avances descritos, evaluar los resultados y comprender los alcances de las conclusiones” (p. 4).

El cuarto criterio fue reiterar algunas de las exigencias y sugerencias de organismos (APA) y autores como Cohen (1988), Cumming y Calin-Jageman (2017) y Maxwell *et al*. (2018) para documentar algunas estadísticas en diseños experimentales, así como explicar los procedimientos detrás de estas. Según la APA (2020):

> Cuando se describen estadísticas inferenciales (e.g., test *t* o *F* asociados con tamaño de efecto e intervalos de confianza), se incluye suficiente información para permitir al lector la comprensión completa de los análisis que se llevaron a cabo. Los datos proveídos, preferentemente en el texto, pero posiblemente en *materiales suplementarios* dependiendo en la magnitud del conjunto de datos, debe de permitir al lector confirmar los reportes básicos de los análisis (e. g., promedios, *SD*, tamaño de la muestra, correlaciones) y debe de *empoderar* al lector interesado para construir estimaciones de tamaños de efectos e intervalos de confianza más allá de los proveídos en el manuscrito *per se* (p. 181).

Ioannidis (2005) muestra que la probabilidad de una afirmación científica de ser cierta/verdadera depende de las siguientes variables:

- Poder.
- Sesgos (manipulación de análisis o en la sección de los resultados, y selección o distorsión de los resultados es una forma típica de sesgos).
- El número de estudios contestando la misma pregunta de investigación (un número reducido hace menos probable que los resultados sean verdaderos).
- Proporción de relaciones verdaderas a no relaciones en las investigaciones (entre mayor sea la proporción de relaciones verdaderas a no relaciones [e. g., 2:1] mayor será la probabilidad de encontrar relaciones verdaderas).

En concreto, Ioannidis (2005) afirma que los "resultados publicados de investigaciones son a veces refutados por evidencia subsecuente, asegurando confusión y decepción" (p. 696). Además, que la mayoría de los hallazgos son falsos y esto es demostrable. Cabe señalar que estas declaraciones fueron emitidas en el contexto de la medicina y habría que ver si son igualmente aplicables en el ámbito de la investigación educativa.

Justificación académica: vacío en la literatura

Durante la primavera del 2020, la búsqueda de la pregunta de la presente investigación produjo cero artículos en Google Académico. En contraparte, en este mismo buscador, se encontraron cuatro publicaciones usando las palabras clave de la pregunta: *estadística*, *investigaciones educativas* y *procesos experimentales*. Tres de ellas eran tesis y la otra, una compilación de investigaciones científicas en ingeniería y educación. Ninguna de ellas cubría la pregunta o los objetivos de la presente. Dados los resultados de Google Académico, se asume un vacío en la literatura.

Aquí se intenta llenar ese vacío con los resultados y análisis de la muestra. Lo anterior justifica esta investigación: cubrir los huecos del conocimiento contribuye a describir, explicar y predecir de mejor manera la realidad (Gall, Gall y Borg, 2007). En efecto, la

ciencia no apunta a otra cosa que "a incrementar nuestro entendimiento de por qué las cosas pasan en la manera que lo hacen" (Carey, 2011, p. 2). Para hacer ciencia, se puede usar la guía de la APA (2020). Otra alternativa para describir estadísticas relacionadas con diseños experimentales fue dada por Nicol y Pexman (2010).

Justificación práctica

Los resultados de este estudio pueden ayudar a seguir ciertos principios para detallar los resultados de investigaciones experimentales (i. e., poder, réplica, acceso a sus bases de datos y reportes de estadísticas). Esto puede contribuir a hacer inferencias mejor fundamentadas en teoría estadística y a crear procesos accesibles, transparentes y con la información necesaria para que otros investigadores puedan hacer réplicas.

Definiciones conceptuales y operacionales

En este trabajo se usa la estadística de la frecuencia o clásica por su gran uso y porque corresponde a la muestra (consultar a Russo [2021] para más información de la teoría de la frecuencia). El test de significancia estadística de la hipótesis nula utiliza la probabilidad calculada $p < \alpha$ (decisión: rechazar la nula, H_0) y $p \geq \alpha$ (decisión: no rechazar la H_0). La p corresponde a los datos y es la probabilidad de encontrar cierta *estadística de la prueba* (e. g., un valor t o F) o una más extrema, cuando la H_0 es cierta. Según Salkind (2007), esta prueba puede ser de utilidad para explorar las siguientes cuestiones:

a) La incertidumbre inherente en los datos empíricos.

b) La naturaleza de las inferencias estadísticas.

c) La estadística de la prueba (e. g., valor de t o F), que representa un resultado de un análisis.

d) La naturaleza de una decisión de rechazar una H_0 en cuanto a un efecto al azar.

Calidad

En la vida cotidiana un usuario compara productos y decide su *calidad*: califica a aquello que está comparando como malo, bueno y excelente (Kotz, 2006; Singh y Khan, 2019). La manera de volver operacional este concepto es someterlo a los cuatro criterios antes descritos.

Otras definiciones

Gall *et al.* (2007) afirmaron que el experimento es el método cuantitativo de investigación más poderoso para establecer las relaciones de causa y efecto entre dos o más variables. Por esta capacidad de establecer la relación de causa y efecto, se seleccionó este diseño para el presente estudio. Este diseño experimental corresponde a los análisis propuestos por la APA y cubiertos por Cohen (1988), Cumming y Calin-Jageman (2017) y Maxwell *et al*. (2018), entre otros. Dicha relación de causa y efecto en un diseño experimental

(DE) puede ser cuestionable por las *amenazas internas y externas* (ver a Gall *et al*. [2007] para ahondar en estas amenazas).

Las características de la muestra aquí seleccionada no empatan del todo con esta definición de *experimento*. Una de las razones fue que en la mayoría de los artículos no se seleccionaron al azar las respectivas muestras poblacionales, así como tampoco se asignaron al azar el *grupo control* (no recibió tratamiento) y el *grupo tratamiento/experimental* (recibió el tratamiento). Por lo tanto, se usó otra definición más apropiada:

> [El] diseño experimental es un plan de los procedimientos para ser seguidos en una experimentación científica para alcanzar conclusiones válidas, con consideraciones de tales factores como selección de participantes, manipulación de variable, recolección de datos y análisis, y minimización de influencias externas (VandenBos, 2015, p. 397).

Un diseño experimental involucra una causa y un efecto: un evento o estado que es traído como resultado de otro (VandenBos, 2015). En otras palabras, la causa o variable independiente es medida en una escala nominal (e. g., sexo: hombre y mujer) y se manipula: se divide la muestra en un grupo control (sin tratamiento o con un placebo; e. g., tutorías) y un grupo tratamiento (e. g., con tutorías). Se puede medir el efecto con la diferencia entre grupos independientes y dentro de grupos dependientes (pre y postest). Para observar el posible efecto entre estas variables, se usa un ensayo estadístico para poner a prueba la H_0 (e. g., test *t* de muestras independientes o dependientes y análisis de la varianza de un factor). La H_0 establece que *no* hay diferencia entre los promedios de las poblaciones, así como tampoco entre las muestras (Russo, 2021; Salkind, 2007).

Ya sea para análisis de grupos independientes o dependientes, el efecto se vuelve operacional al obtener su magnitud: “Tamaño de efecto, que es una o varias mediciones de la magnitud o el significado de la relación entre dos variables. Seguido, los tamaños de efecto son interpretados como indicativos de la significancia práctica de un resultado de investigación” (VandenBos, 2015, p. 352). Para la presente investigación, se tomó en cuenta el *efecto principal.* VandenBos (2015) lo definió de la siguiente manera: “El efecto total consistente de una sola variable independiente sobre una variable dependiente sobre todas las demás variables independientes en un diseño experimental. Este es distinto de, pero puede ser oscurecido por, un efecto de interacción entre variables” (p. 617).

En su libro seminal, Cohen (1988) dio varios tamaños para clasificar los efectos: pequeño, mediano y grande; pero advirtió que estos tamaños solo deben de ser usados cuando no existe un contexto para ser interpretados. Para ahondar en otros tamaños de efecto, se recomienda a Hancock, Stapleton y Mueller (2019) y a Sakai (2018) para experimentos en laboratorio. Está más allá de los objetivos de este presente manuscrito el evaluar el tamaño del efecto de la muestra.

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

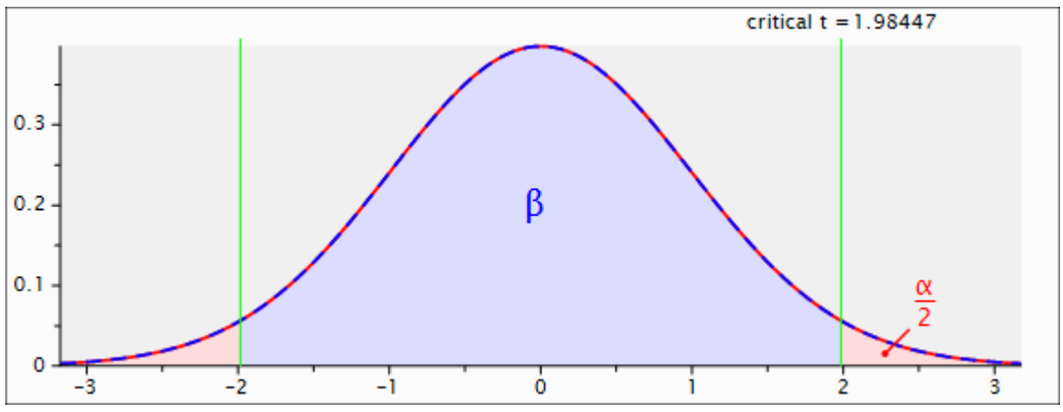
Significancia estadística

Las figuras 1-6 contienen una serie de ejemplos para explicar los conceptos de *significancia estadística*, *hipótesis nula* (H_0) y *alternativa* (H_A), *α* (nivel de significancia; error tipo I [ETI]), beta (β; error tipo II [ETII]), *poder*, *muestras*, *poblaciones* y *error de muestreo*. La significancia estadística es "el grado al cual un resultado de una investigación no puede ser atribuido razonablemente a la intervención del azar u otros factores aleatorios" (VandenBos, 2015, p. 1026). Con la prueba de significancia estadística se pone a prueba a la H_0: el fenómeno no existe (Fisher, 1949).

Ahondando al respecto, Ellenberg (2015) explicó: "Un resultado estadísticamente significativo nos da una pista, sugiriendo una tierra prometida para enfocar la energía investigativa" (p. 156). Más detalladamente, este autor mencionó que una prueba de significancia estadística funge como si fuera un detective, no un juez *per se*. En otras palabras, un $p < \alpha$ es el inicio de algún resultado prometedor con una serie de réplicas (Cumming y Calin-Jageman, 2017), pero no un fin en sí mismo. Si se rechaza la H_0, se concluye que el fenómeno existe (i. e., evento observable, según VandenBos [2015]).

Entonces, el valor *p* indica qué tan sorprendentes son los resultados de los análisis bajo el supuesto de que no hay efecto (figura 1; el área sombreada de azul es el área de no rechazo de la nula: 1 – α: comúnmente es 1 – 0.05 = 0.95). Esto quiere decir que si se comparan los promedios de un examen de matemáticas de dos grupos probablemente no resulten iguales. Si la diferencia entre estos promedios no es muy grande, $p \geq \alpha$, estos resultados *no* son sorprendentes y la diferencia se debió al azar. Por otro lado, si $p < \alpha$, los resultados serían sorprendentes, bajo el supuesto de que no existe diferencia entre las poblaciones: i. e., cabe la posibilidad de algún efecto. En síntesis, cuando la nula es verdadera (figura 1; no hay un efecto verdadero: las distribuciones de ambos grupos se superponen perfectamente), la probabilidad de encontrar resultados estadísticamente significativos es igual al α.

Figura 1. Dos distribuciones normales perfectamente superpuestas



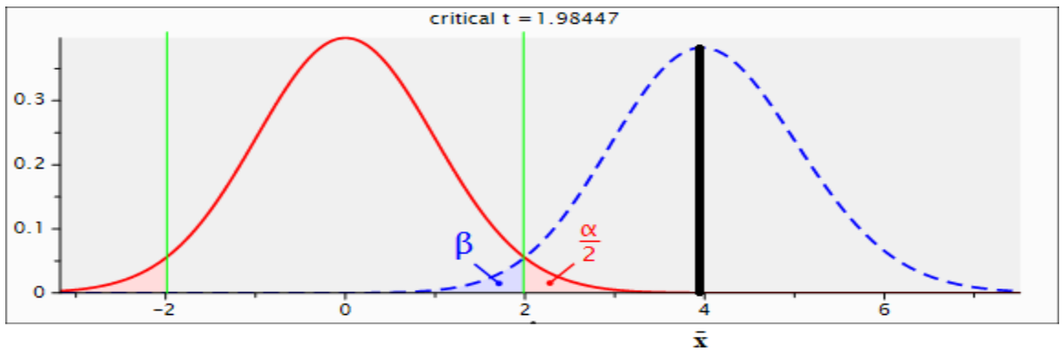
Nota: Modelo de rechazo de la H_0, así como una H_0 cierta. Se aplicó una prueba *t* de grupos independientes dado: poder = 0.05; α = 0.05/2; β = 0.95; *t* crítico = 1.98, *df* = 98 y $p > \alpha$.

Fuente: Elaboración Propia

En contraparte, si la nula es falsa (figura 2; hay un efecto verdadero) la probabilidad de encontrar resultados estadísticamente significativos es igual al poder (e. g., 0.80 u 80 % recomendado como mínimo por Cohen [1988] y Cumming y Calin-Jageman [2017]). De

hecho, Cohen (1988) simplemente lo definió de la siguiente forma: "El poder de una prueba estadística es la probabilidad que esta prueba dará resultados significativos" (p. 1). En pocas palabras, el poder es la probabilidad de obtener resultados significativos cuando la nula es falsa.

Figura 2. Dos distribuciones normales separadas



Nota: *d* cambia a 0.80. Esto hace que el poder sea 0.97 y $p < \alpha$: se rechaza la nula. La $\bar{x}$ significa el promedio del grupo de tratamiento.

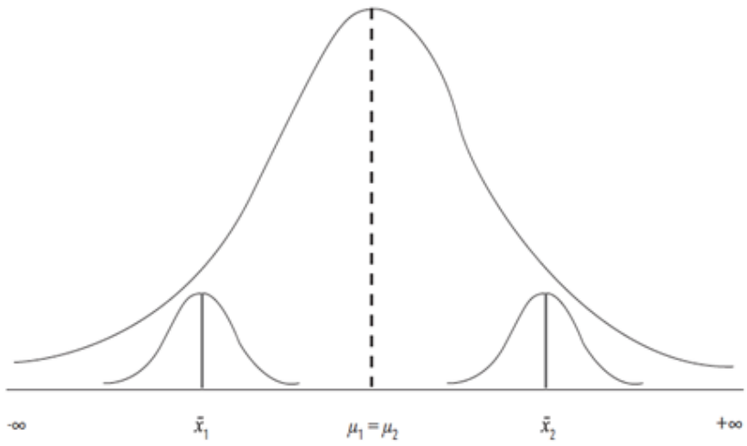
Fuente: Elaboración propia

Suponiendo que una investigadora lleva a cabo un diseño experimental (Estudio I), toma una muestra ($n = 100$) al azar de la población (hay que recordar que de las poblaciones se obtienen parámetros: e. g., μ = promedio de la población, σ = desviación estándar de la población y $\delta = d$ de Cohen de las poblaciones) y de las muestras obtiene estadísticas ($\bar{x}$ = promedio de la muestra, SD = desviación estándar de la muestra y $d = d$ de Cohen de las muestras). Divide la muestra al azar en dos para tener un grupo control (50) y un grupo tratamiento (50); implementa una prueba *t* de muestras independientes con un $\alpha = 0.05$ dividido en dos colas $\alpha / 2$: α = también representa el error tipo I: probabilidad de rechazar una H_0 cuando es cierta. Asume que ambas muestras vienen de diferentes distribuciones de poblaciones (figura 1), pero estas son idénticas porque sus parámetros son iguales (H_0 es cierta): i. e., $\mu_1 = \mu_2$ y $\sigma_1 = \sigma_2$, y, por lo tanto, $\delta = (\mu_1 - \mu_2) / \sigma = 0$. El área de las dos distribuciones sobrepuestas es igual a uno o 100 %. Por el otro lado, está el valor beta (probabilidad de no rechazar una H_0 cuando es falsa: 1 – poder): en este caso resulta que $\beta = 1 - 0.05 = 0.95$ (figura 1), y el poder es $1 - \beta$ ($1 - 95 = 0.05$), al igual que el $\alpha = 0.05$.

Algo que puede pasar por error de muestreo (error de diferencia entre los parámetros de la población y las estadísticas de la muestra) es mostrado en la figura 3: H_0 es cierta, pero las dos muestras ($\bar{x}_1 \neq \bar{x}_2$, $d \neq 0$; $p < \alpha$). Basado en las estadísticas de estas muestras, se rechaza la H_0 erróneamente (i. e., error tipo I).

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

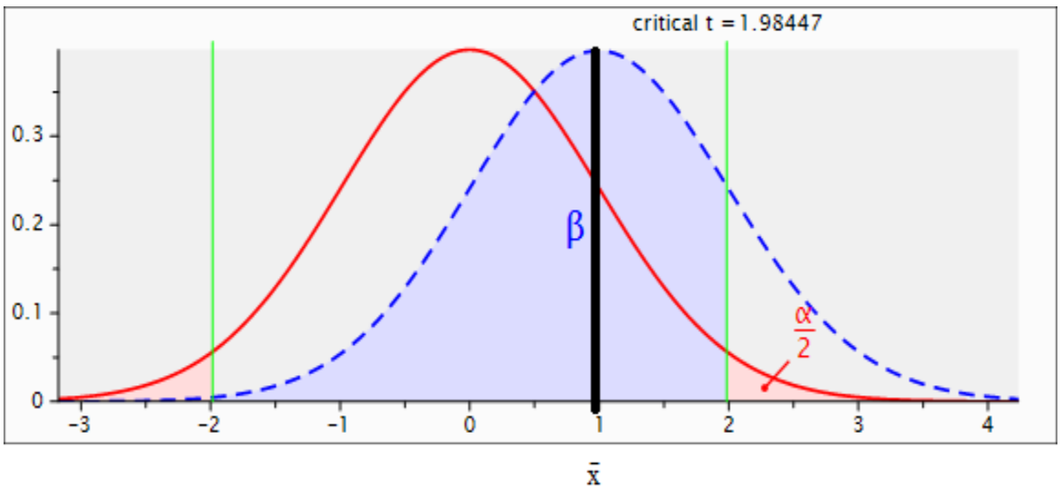
Figura 3. H_0 es cierta con error tipo I



Fuente: Ponce (2019)

Volviendo al ejemplo de la investigadora, ella aplica un tratamiento a un grupo y nada al otro, y se parte de una H_0 cierta. Como resultado se obtiene la figura 4, donde se observa cómo las dos muestras representadas por las dos distribuciones normales (la curva sólida representa a la distribución de la muestra del grupo control y la curva punteada a la del grupo tratamiento) no se sobreponen ($\bar{x}$ control ≠ $\bar{x}$ tratamiento): el tratamiento causó que las distribuciones se separaran con un d = 0.20 y el poder ≈ 0.17 (β ≈ 1 – .17 ≈ .83), pero el $\bar{x}$ tratamiento no cayó en el área de rechazo de la nula (área sombreada de las colas: $p \geq \alpha$). En este caso, no se puede saber si se debió a la falta de un efecto verdadero del tratamiento o a no tener una muestra lo suficientemente grande con el poder para detectarlo. Sin embargo, para sacar conclusiones sobre la efectividad el efecto práctico del tratamiento habría que ver la literatura para sacar alguna conclusión (Cohen, 1988; Cumming y Calin-Jageman, 2017).

Figura 4. Dos distribuciones separadas por 0.20 de SD

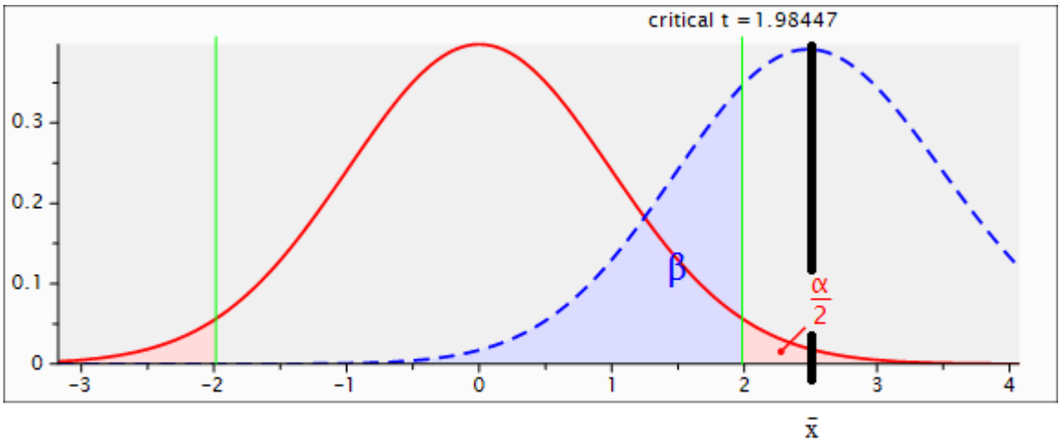


Nota: d cambió a 0.20; poder ≈ 0.17; α = 0.05/2; β ≈ 0.83; t crítico = 1.98, df = 98. La curva punteada representa, junto con la curva sólida, la hipótesis alternativa (H_A: $\mu_1 \neq \mu_2$).

Fuente: Elaboración propia

Otro investigador replica el Estudio I (figura 5); encuentra un d = 0.50 con una $p < \alpha$, β ≈ 0.30, y el promedio del grupo tratamiento cae en el área de rechazo de la nula. Sin embargo, el poder (≈ 0.70) no llega a 0.80, así que se recomendaría subirlo, lo que incrementaría el tamaño de la muestra.

Figura 5. Dos distribuciones separadas por 0.50 de SD

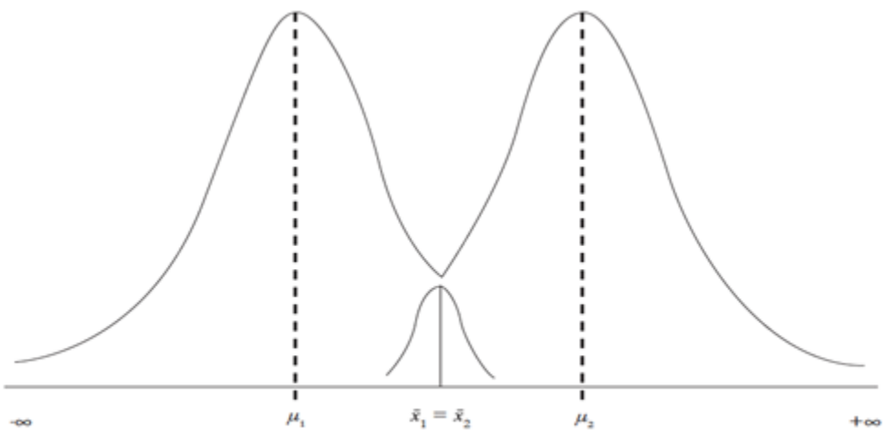


Nota: *d* = 0.50; Poder ≈ .70; α = .05/2; β ≈.30; *t* crítico = 1.98, *df* = 98y *p* < alfa.

Fuente: Elaboración Propia

Un tercer investigador lleva a cabo otra replicación del Estudio I. Su tratamiento tiene un *d* = 0.80 con $p < \alpha$, β ≈ 0.03, y un poder ≈ 0.97, así que contó con suficiente poder: este sería un resultado sorprendente para seguir investigando (figura 2). Para terminar, la figura 6 plantea que las muestras ($\bar{x}_1 = \bar{x}_2$) cuando la H_0 fue falsa ($\mu_1 \neq \mu_2$), así que se comete un error tipo II al no rechazar una H_0 falsa.

Figura 6. H_0 falsa con error tipo II



Fuente: Ponce (2019)

Implicaciones del error I y II

Una implicación del error tipo I es obtener un falso positivo [FP] (e. g., identificar a una estudiante con aptitudes sobresalientes cuando en realidad no las tiene). Asimismo, una implicación del error tipo II es encontrar un falso negativo [FN] (e. g., un estudiante no califica para educación especial cuando debió de ser atendido). Esta probabilidad de obtener un FP se puede calcular cuando se estima la probabilidad de una H_0 de ser cierta o falsa, dado un α. Si una H_0 tiene la probabilidad de ser cierta de 50 % y se selecciona un α = 0.05 o 5 %, la probabilidad de un FP es 50 % × 5 % = 2.5 % (i. e., 0.025). La tabla 1 muestra las otras probabilidades de obtener FP, así como un verdadero positivo [VP] (e. g., un estudiante con

aptitudes sobresalientes que es identificado como tal) y un verdadero negativo [VN] (un estudiante sin aptitudes sobresalientes que es identificado como que no las tiene).

Tabla 1. H_0 cierta o falsa

Resultado	H_0 cierta: 50 %	H_o falsa/H_A cierta: 50 %
$p < \alpha$	FP (α) 0.05 × 0.50 = 0.025	*VP* *(1 - β)* *0.80 × 0.50 = 0.40*
$p \geq \alpha$	*VN* *(1 - α)* *0.95 × 0.50 = 0.475*	FN (β) 0.20 × 0.50 = 0.10

Fuente: Lakens (s. f.)

Relación entre el poder y otras variables

El poder puede ser incrementado al aumentar el tamaño de la muestra (Cohen, 1988; Cumming y Calin-Jageman, 2017). Las figuras 7-11 muestran las relaciones entre las siguientes variables: el poder y FP, FN, VN y VP cuando las H_0 tiene cierta probabilidad de ser cierta o falsa. Los siguientes cinco ejemplos muestran estas relaciones entre estas variables.

1) El poder tiene una relación inversamente proporcional al coeficiente beta (β): i. e., cuando el poder aumenta 1 %, el beta (FN) disminuye 1 % (figura 7). El poder no tiene efecto en el α, pero al aumentar α sí incrementa el poder (ver a Cohen, 1988; Ponce, 2019).

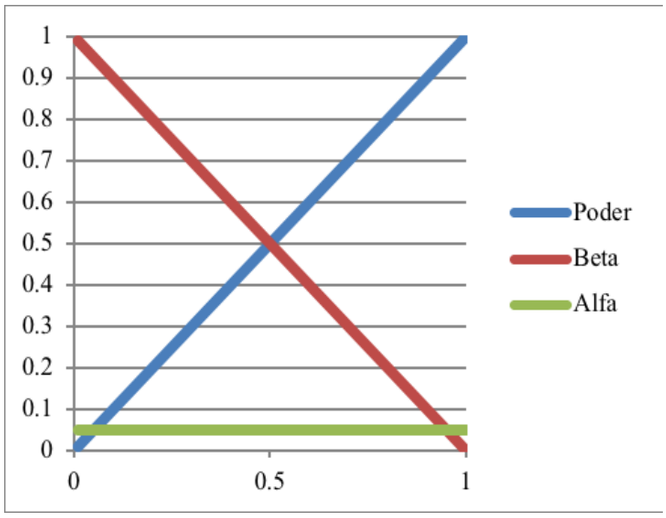
2) Dado la H_0 con 50 % de ser cierta, aunque el poder aumente (figura 8), un FP dado $p < \alpha$ se mantiene fijo al igual que el VN dado un $p \geq \alpha$ (ver tabla 1 para los cálculos matemáticos de la probabilidad).

3) Cuando la H_0 tiene 50 % de ser falsa (figura 9), al aumentar el poder, aumenta la probabilidad de un VP (cuando $p < \alpha$) y la probabilidad de un FN disminuye (cuando $p \geq \alpha$).

4) Cuando la probabilidad de la H_0 de ser verdadera incrementa (figura 10), teniendo el poder constante a 80 %, la probabilidad de un FP aumenta, pero su pendiente es menor a la otra probabilidad que también aumenta, un VN. Las probabilidades que disminuyen son las de VP y de FN (este último tiene una pendiente menor).

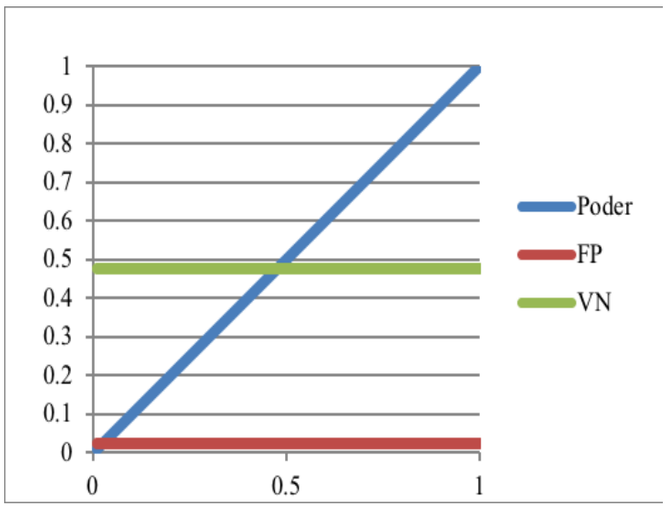
5) Cuando la probabilidad de que una H_0 sea falsa aumenta (figura 11), manteniendo el poder a 80 %, la probabilidad de un FP y un VN disminuyen (este último tiene una pendiente más pronunciada). Por el otro lado, la probabilidad de un VP (mayor pendiente) y un FN aumentan.

Figura 7. Relación entre el poder y el beta



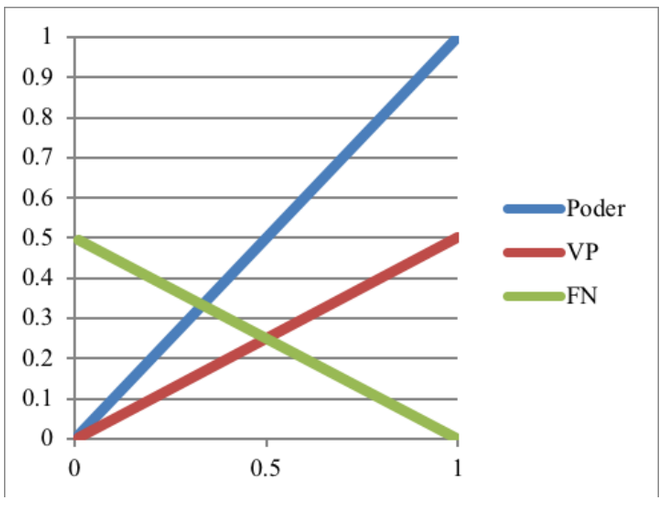
Fuente: Elaboración propia

Figura 8. La H_0 tiene 50 % de ser cierta



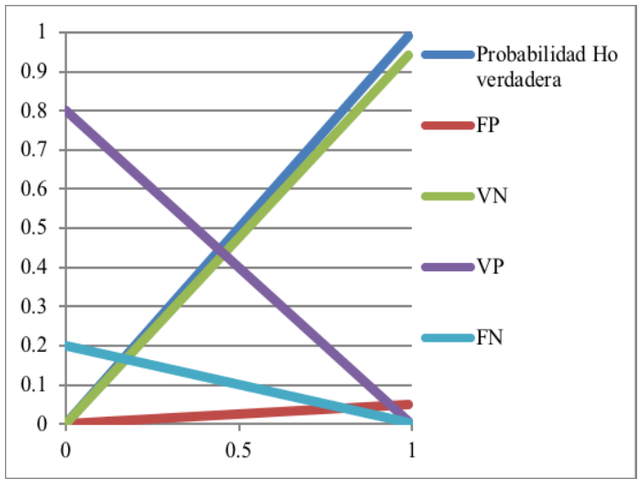
Fuente: Elaboración propia

Figura 9. La H_0 tiene 50 % de ser falsa



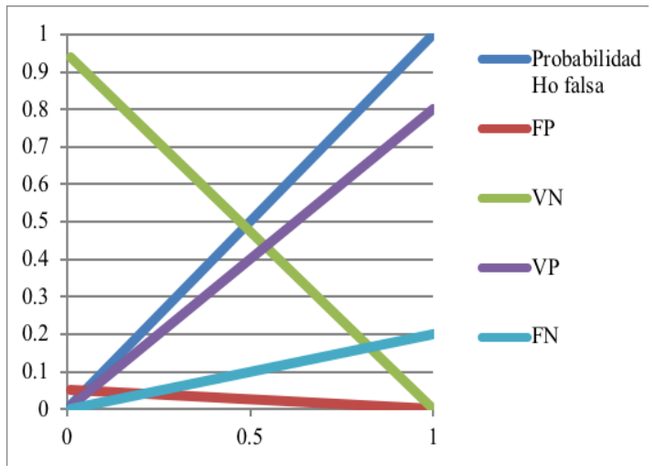
Fuente: Elaboración propia

Figura 10. Probabilidad de un H_0 verdadera



Fuente: Elaboración Propia

Figura 11. Probabilidad de una H_0 sea falsa



Fuente: Elaboración propia

En el supuesto que la H_0 es cierta, esto quiere decir que la población de la que vino el grupo control y el grupo tratamiento son iguales. Es decir, vienen de la misma población porque ambas distribuciones se superponen perfectamente la una a la otra. Por ello, si hay una diferencia entre ambas al final del experimento, se deberá al tratamiento que causó esta discrepancia (sin considerar las amenazas internas y externas al experimento). Si la H_0 es falsa (hay una diferencia entre los promedios de estas: efecto cierto; e. g., algún tratamiento que mejoró el promedio), entonces la probabilidad de que un segundo estudio resulte en un efecto verdadero es igual al poder que se tenga en este, otras cosas siendo iguales (Lakens, s. f.).

Método

Se usó una metodología cuantitativa, exploratoria, con estadística descriptiva e inferencial con el fin de evaluar la calidad de 34 artículos. Para ello, cuatro criterios fueron utilizados:

1) Haber discutido, calculado y obtenido el suficiente *poder* para rechazar una hipótesis nula falsa.

2) Observar si el estudio en cuestión es parte de una serie de *réplicas*.

3) Ver si se dio *acceso* a las bases de datos para potenciales réplicas.

4) Reportar estadísticas sugeridas por la APA (2001, 2020) desde el 2001 hasta el presente día.

Los artículos de la muestra fueron publicados en una serie de revistas arbitradas en el ranking de la Red Iberoamericana de Innovación y Conocimiento Científico (Redib). Se seleccionó a la Redib porque la audiencia potencial para este estudio podría encontrarse entre investigadores que han publicado en las revistas de la muestra o las usan para sus trabajos. Además, la Redib contiene enlaces a revistas de libre acceso, lo cual cumple con el principio de accesibilidad al conocimiento que se busca también en este manuscrito.

En resumen, el método consistió en definir ciertas bases para seleccionar una serie de artículos de revistas electrónicas (ver la siguiente sección). Luego, se leyeron estos para capturar información y estadísticas referentes a una serie de criterios para analizar su calidad (el instrumento consistió en una tabla para capturar información, ver tabla 2). Parte de los análisis a las estadísticas de los artículos fue descriptiva (tabla 3) y otra parte fue con estadística inferencial (tablas 4 y 5). En los siguientes párrafos del método se dan más detalles de los pasos anteriores, información y estadísticas capturadas, así como análisis efectuados.

El instrumento de análisis fue a través de un documento de Excel (tabla 2), donde se capturaron los diversos elementos seleccionados, para el posterior análisis en el *software* estadístico SPSS versión 25.

Tabla 2. Resultados de los cuatro criterios para evaluar a la muestra de 34 artículos

Respuesta	*1)* Poder	*2)* Réplica	*3) Acceso*	*4)* $\bar{x}$	*4) SD*	*4)* IC	*4)* TE	*4)* TA	*4)* Est.	*4) df*	*4) p*
Sí				Sí*	Sí*	Sí*	Sí*	Sí*	Sí*	Sí*	Sí*
No	No	No	No								
¿Cuál fue?				**	**	**	**	**	**	**	**

Nota: Sí = Sí reportó el valor; No = No reportó el valor; ¿Cuál fue? = El coeficiente reportado es el valor $\bar{x}$ = promedio; SD = desviación estándar; IC = Intervalo de Confianza; TE = Tamaño del efecto; Est. = Estadística *t* y *F*; *df* = grados de libertad, y *p* = probabilidad calculada. Sí* = Algunos sí reportaron. ** = Los valores están en la tabla 3.

Fuente: Elaboración propia

Muestra

Los artículos de la muestra (*n* = 34) del presente estudio fueron obtenidos de revistas (20) que aparecieron en el ranking del 2018 de la Redib: área temática (ciencias sociales y humanidades); materias (educación e investigación educativa), y país (todos los países). Solo se incluyeron artículos en español e inglés por ser leíbles por los presentes autores. Las palabras clave fueron: *experimento* (21 artículos en español) y *experiment* (13 en inglés). La población de esta muestra fueron 92 revistas y 502 artículos. Las bases para seleccionar los artículos fueron:

- Haber llevado a cabo algún *diseño experimental* en investigación educativa: i. e., comparar grupos cuando uno de ellos recibió algún tratamiento que lo diferenció del otro, así como un tratamiento recibido entre el pre y el postest.
- Haber usado un análisis *paramétrico* (i. e., una prueba *t* o *F*) porque es recomendable para inferir causas y efectos (Maxwell *et al*., 2018). Algunos artículos también contenían algunos no paramétricos.

El intervalo de tiempo de publicación de los artículos fue del 2004-2019: específicamente 23.4 % cubrió el intervalo del 2004-2012 y 76.5 % (2013-2019). En los artículos, se llevaron a cabo tratamientos de aprendizaje del idioma inglés y efecto de diferentes formatos de exámenes, entre otros.

Se buscó en los artículos de la muestra información relacionada con estos cuatro criterios, pero ninguno de los artículos de la muestra cumplió con ninguno de los tres primeros criterios (ver tabla 2 y la sección de resultados para más detalles).

Para el criterio cuatro, se extrajeron los siguientes datos de la muestra de artículos para ser analizados con estadísticas inferenciales (tabla 3):

- Tamaño de la muestra (*n*): todos los autores describieron el tamaño de su muestra de estudio cumpliendo con el APA (2001, 2020).
- Promedios, SD, tipo de análisis, estadística de la prueba (e. g., valor calculado *t* y *F*), grados de libertad (*df*) y *p*. Hubo cierta variación en reportar estas estadísticas.
- Intervalos de confianza (IC) y tamaño de efecto. Al contrario, otras estadísticas fueron las menos especificadas: incumpliendo con el APA (2001, 2020).

Aunque no era parte de los criterios de calidad, la sección B de la tabla 3 contempla un principio de la epistemología de la ciencia (modelo parsimonioso): un modelo más simple con más poder para explicar y con menos supuestos y variables es preferible a otro más complicado (Sarkar y Pfeifer, 2006). En el contexto del presente trabajo, varios de las y los autores compararon tres promedios de tres grupos independientes con una prueba *t*, haciendo lo siguiente: el grupo dos con el dos, el uno con el tres y el dos con el tres (inflación). Para corregir esta inflación, se debe de ajustar el α usando la *corrección de Bonferroni*: α / número de pruebas (α / *n*) (Armstrong, 2014). Usando el ejemplo anterior de tres pruebas *t*, el α debió de ser ajustado a 0.05 / 3 = 0.016. Además, se puede usar un modelo más parsimonioso para el ejemplo anterior: el análisis de la varianza (Anova) de un factor que necesita una comparación entre los tres grupos. La advertencia es que, si solo se está interesado en una prueba, no es necesario ajustar el α. Con base en los artículos de la muestra que usaron varias variables independientes y dependientes en varias ocasiones, se recomienda usar el análisis multivariante de la varianza (Manova) (Maxwell *et al*., 2018).

La sección C de la tabla 3 enseña las medidas de tendencia central del tamaño de las muestras de los artículos. Por ejemplo, el promedio del tamaño de los artículos fue de 75.12 participantes. También, se muestran las medidas de dispersión de los tamaños de la muestra. Un ejemplo es la desviación estándar, que fue de 46.36 participantes; también se muestra un intervalo de confianza de 95 %, valor mínimo y máximo. Además, se cuantificaron las medidas de tendencia central y variación del número de análisis de los artículos de la muestra. Por ejemplo, el promedio de número de análisis de los artículos fue de 14.94 para los paramétricos (SD = 26.21) y para los no-paramétricos 1.67 con una SD = 1.19.

Dada la carencia de la descripción del poder en los estudios, se decidió estimar este factor con parte de la información allí incluida (aunque en la mayoría de los casos la información no fue suficiente porque faltaron tamaños de efecto para calcular el poder). Se usó G*Power (*software* gratuito que puede ayudar a calcular *n a priori* para tener el suficiente poder, así como *a posteriori*; creado y documentado por Faul, Erdfelder, Buchner y Lang, 2009). En la sección C de la tabla 3 se muestran las medidas de tendencia central y la variación del poder que fue estimado para el presente manuscrito (e. g., promedio del poder = 0.74 con una SD = 0.27).

Un análisis de poder es importante porque permite a las y los investigadores planificar qué recursos se necesitarán para inscribirse o seleccionar la cantidad deseada de individuos para el estudio (APA, 2020). A manera de describir los datos, se calculó la posible distribución normal de los coeficientes de poder. Las estimaciones del poder no pasaron la prueba de Kolmogorov-Smirnov para distribuciones normales ($p = 0.001 < \alpha = 0.05$). Otra evidencia en contra de la normalidad de los datos fue la diferencia entre las estadísticas de tendencia central que *no* coinciden: promedio (0.74), mediana (0.83) y moda (0.99): el promedio y el IC son afectados por una distribución anormal. Sin embargo, la curtosis (-0.679) y el sesgo (-0.712) estuvieron entre ± 2, lo que indica una posible distribución normal. Dadas estas evidencias contradictorias, se optó por un análisis con porcentajes: los valores del poder estimados por encima de 80 % fueron de 58.8 % y por debajo de este de 41.2 %. Es decir, la mayoría de los artículos de la muestra hubieran cumplido con el poder estadístico de 80 %, otras cosas siendo iguales. Sin embargo, era el trabajo de estas y estos autores el mostrar que tenían el suficiente poder para argumentar algún efecto de sus diseños experimentales.

Tabla 3. Datos descriptivos de artículos de la muestra

Sección A	Número Encontrado	% Encontrado	Número No encontrado	% No encontrado
a) Promedio	30	88.24	4	11.76
b) SD	26	76.47	8	23.53
c) IC	6	17.65	28	82.35
d) Tamaño de efecto	12	35.29	22	64.71
**e)* Tipo de análisis	32	94.1	2	5.9
f) Estadística de la prueba (valor *t* o *F*)	28	82.4	6	17.6
g) df	22	64.71	12	35.29
h) p	30	88.24	4	11.76
Sección B	Cometido	%	No cometido	%
i) Inflación del error tipo I	33	97.06	1	2.94
Sección C	*N*	Test paramétricos elaborados	Test no paramétricos elaborados	Poder estadístico estimado*
Promedio	75.12	14.94	1.67	0.74 $CI_{95\ \%}$ [0.649, 0.831]
Mediana	64	4.5	0	0.83 $CI_{97.5\ \%}$ [0.53, 0.99]
Moda	48	2	0	0.99
SD	46.36	26.21	1.19	0.27
$IC_{95\ \%}$	[59.30, 90.94]	[13.4, 16.4]	[1.27, 2.07]	
Valor mínimo	10	1	0	0.08
Valor máximo	205	120	39	1

* 38.2 % fue una prueba *t*; 26.5 % fue test *F*; 17.6 % fue una combinación una prueba *t* y *F*, y 11.8 % fue otra combinación entre test paramétrico y no paramétrico (e. g., test de ji al cuadrado). Ahora bien, 41.2 % de los valores estimados del poder estadístico estuvo por debajo de 0.80 y 58.8 %. Las estadísticas de los artículos y sus correspondientes revistas pueden ser revisadas en: https://cos.io.

Fuente: Elaboración propia

Análisis estadístico inferencial

Para los análisis inferenciales, se utilizaron dos tipos de análisis con la prueba de ji al cuadrado (χ^2): *a)* de bondad de ajuste y *b)* de independencia. Esto para ver, con el uso de SPSS versión 25, si existía una diferencia entre las frecuencias esperadas y las observadas. Ambas pruebas fueron tomadas de Berenson, Levine, Szabat y Stephan (2019), Greenwood y Nikulin (1996), Hinkle, Wiersma y Jurs (2003) y Kohler (2020).

La prueba de bondad de ajuste involucra una sola muestra (frecuencia observada) que se compara con una frecuencia esperada, que está basada en cierta expectativa. En este caso, la expectativa era que 85 % de los artículos cumpliera con reportar alguna de las estadísticas de la tabla 4: (i. e., promedio, SD, IC, tamaño del efecto, tipo de análisis, estadística de prueba, *df* y *p*). Por lo tanto, se esperaba que 15 % no reportara estas estadísticas antes mencionadas. La expectativa se tomó de uno de los supuestos de χ^2, a saber: es necesario tener por lo menos cinco casos en la *frecuencia esperada* para poder usar la probabilidad calculada que se da para la estadística de la prueba χ^2 (Berenson *et al*., 2019). Aquí 5 de 34 artículos representan 15 % (la frecuencia esperada). Las hipótesis de la prueba de bondad de ajuste son las siguientes:

- Hipótesis nula (H_0): la frecuencia observada = la frecuencia esperada.
- Hipótesis alternativa (H_A): la frecuencia observada ≠ la frecuencia esperada.

Ya que se hicieron ocho comparaciones, el alfa (α) debe ser ajustado para evitar inflación del error tipo I con la ya antes mencionada corrección de Bonferroni. Así, tras dividir el alfa (i. e., 0.05) entre el número de pruebas, obtuvimos como resultado que el coeficiente alfa ajustado fue 0.05 / 8 = 0.0063. Por lo tanto, el criterio para rechazar la hipótesis nula fue: $\alpha_{\text{ajustado}} = 0.0063$.

Tabla 4. Tabla cruzada de bondad de ajuste 2 × 2

Reportes	Expectativa	Estadística observada*
Reportó	29	Sí reportó (frecuencia)
No reportó	5	No reportó (frecuencia)

Nota: * promedio, SD, IC, tamaño del efecto, tipo de análisis, estadística de prueba, *df* y *p*.

Fuente: Elaboración propia

Otro aspecto de la prueba de bondad de ajuste es el tamaño del efecto: *W* de Cohen, que mide la discrepancia entre los pares de proporciones en las celdas (Cohen, 1988). Entre más grande sea la diferencia entre la frecuencia esperada y la observada mayor será el tamaño del efecto. Los tres diferentes tamaños son: $W = 0.10$ (pequeño), $W = 0.30$ (mediano) y $W = 0.50$ (grande), cuando no se tiene alguna referencia en la literatura.

Asimismo, la prueba de independencia de χ^2 involucra dos variables para ver qué tanto se relacionan. Si no se relacionan: frecuencia esperada = frecuencia observada: $p \geq \alpha$ y se concluye que son independientes. Por el contrario, si se relacionan: frecuencia esperada ≠ frecuencia observada: $p < \alpha$ y se concluye que son dependientes. En esta ocasión, las hipótesis fueron:

- Hipótesis nula (H_0): las dos variables categóricas son independientes (la frecuencia observada = la frecuencia esperada).
- Hipótesis alternativa (H_A): las dos variables categóricas son dependientes (la frecuencia observada ≠ la frecuencia esperada)

El criterio para rechazar la hipótesis nula fue un tradicional $\alpha = 0.05$. dado que es solo una prueba ómnibus de dos variables (2 × 8): reportó (sí o no) y estadística (ocho niveles: promedio, SD, IC, tamaño del efecto, tipo de análisis, estadística de prueba, *df* y *p*). El tamaño de efecto usado en una prueba de independencia es el *V* de Cramer (Akoglu, 2018): i. e., con un $df = 1$, un $V = 0.10$ (pequeño), $V = 0.30$ (mediano) y $V = 0.50$ (grande).

Tabla 5. Tabla cruzada de prueba de independencia (2 × 8)

Estadística	Sí reportó	No reportó
Promedio	30	4
SD	26	8
IC	6	28
Tamaño de efecto	12	22
Tipo de análisis	32	2
Estadística de la prueba (valor *t* o *F*)	28	6
Df	22	12
P	30	4

Nota: todas las pruebas tuvieron un grado de libertad ($df = 1$)

Fuente: Elaboración propia

Por otro lado, una prueba ómnibus indica que $p \geq \alpha$ o $p < \alpha$, pero no indica exactamente dónde puede estribar la diferencia entre la frecuencia esperada y la observada. Para identificarla, se lleva a cabo un análisis *post hoc* (descrito por Beasley y Schumacker [1995]). En concreto, se calculan los *residuales ajustados estandarizados* en SPSS 25. Manualmente, estos residuales ajustados estandarizados se elevan al cuadrado para obtener valores de $\chi^2_{\text{calculado}}$ por cada nivel (en este caso, son ocho niveles por dos = ocho). Con estos valores de $\chi^2_{\text{calculado}}$, se calcula la probabilidad calculada (*p*) con una función de SPSS (Beasley y Schumacker, 1995). Para esta prueba *post hoc* y para evitar la inflación del error tipo I, es necesario usar la corrección de Bonferroni, el valor *p* a las 14 pruebas: 0.05 / 16 = 0.0031.

Resultados

Para los resultados descriptivos, tres criterios *no* se cumplieron porque en 100 % de las publicaciones no se mostró el *poder* en los análisis, no fueron parte de *réplicas* en otros estudios y *no* hubo *accesibilidad* a bases de datos (ver tabla 2). Una de las expectativas personales era encontrar que la mayoría de los autores y autoras mencionaran en sus artículos el poder estadístico de alguna forma. Con respecto al elemento réplica, se esperaba que algunos fueran repetición de otros. Asimismo, se esperaba que aproximadamente 76.5 % de la muestra hubiera dado acceso a sus bases de datos, porque solo este porcentaje corresponde

al 2013 cuando se creó el Center for Open Science: esta no ha sido la única posibilidad para almacenar bases de datos, pero se tomó como expectativa.

Para los resultados inferenciales de las pruebas de χ^2 de bondad de ajuste, se encontró que tres de las ocho pruebas (con un criterio para rechazar la hipótesis nula de $\alpha_{ajustado} = 0.0063$) tuvieron una diferencia estadísticamente significativa con respecto a la expectativa de que sí reportaron = 29 y no reportaron = 5 (tabla 5). Estas fueron el IC (sí reportaron = 6 y no reportaron = 28), el tamaño del efecto (sí = 12 y no = 22) y los grados de libertad (*df*; sí = 22 y no = 12). Por lo tanto, se rechaza la hipótesis nula en el caso de estas tres estadísticas y la evidencia apoya la alternativa de que las frecuencias observadas son diferentes a las esperadas y probablemente esto no es al azar sino allí hay un efecto. El tamaño del efecto *W* fue de 2.47, 1.81 y 0.74 para estas tres estadísticas respectivamente (tabla 6) y esto las ubica en un efecto grande, según Cohen (1988), que significa que existe una diferencia grande entre las frecuencias. Sería complejo especular sobre la razón por la que se omitió el reporte de estas estadísticas dado que desde el APA (2001) hasta la fecha ha insistido en reportarlas. Se recomienda a Cumming y Calin-Jageman (2017), así como al APA (2020) para ver detalles e implicaciones de estas estadísticas.

Tabla 6. Resultados de la prueba de bondad de ajuste

Estadística	χ^2 calculada	*df* (de los análisis de χ^2)	*p*	*W*
Promedio	0.279	1	0.597	0.12
SD	1.940	1	0.164	0.31
IC	120.971	1	< 0.00001*	2.47
Tamaño de efecto	65.885	1	< 0.00001*	1.81
Tipo de análisis	2.217	1	0.137	0.33
Estadística de la prueba (valor *t* o *F*)	0.187	1	0.666	0.10
Df	10.983	1	0.001*	0.74
P	0.279	1	0.597	0.12

Nota: la frecuencia esperada fue de 29 (sí reportó) y cinco (no reportó). *
=Estadísticamente significativo con un $\alpha_{ajustado} = 0.0063$.

Fuente: Elaboración propia

Para los resultados inferenciales de la prueba ómnibus de χ^2 de independencia, se encontró que $\chi^2_{calculado} = 84.817$, $df = 7$, $p < 0.00001$ y *V* de Cramer = 0.558 (con un criterio para rechazar la hipótesis nula de $\alpha = .05$). Por lo tanto, se rechaza la hipótesis nula y se concluye también que existe una dependencia entre las variables. Asimismo, el tamaño del efecto *V* fue de 0.558 (grande) (Akoglu, 2018; Cohen, 1988), que significa que existe una relación grande entre las variables. Sin embargo, como es una prueba ómnibus, no se sabe dónde radica la diferencia entre las frecuencias esperadas y los observadas, así que se usó la prueba *post hoc* de χ^2 de Beasley y Schumacker (1995) con un $\alpha_{ajustado} = 0.0031$. Esta última prueba mostró que la diferencia estadísticamente significativa radicó en el IC (sí reportaron = 6 y no reportaron = 28), tamaño del efecto (sí = 12 y no = 22) y el tipo de análisis (sí = 32

y no = 2). De nueva cuenta, el IC y el tamaño del efecto vuelven a aparecer en la prueba de independencia, donde sus frecuencias observadas son diferentes a las esperadas, son menos reportadas de lo que se esperaría. En contraparte, el tipo de análisis también tuvo una diferencia estadísticamente significativa, pero contrario a estas dos últimas estadísticas, porque fue la estadística más reportada de las ocho (ver tabla 5). En palabras llanas, el reportar un valor de una estadística depende del tipo de estadística de la que se esté hablando. En este caso, el IC y el tamaño del efecto fueron infrautilizados y el tipo de análisis fue sobreutilizado. De nuevo, sería complejo especular sobre la razón que llevó a omitir el reporte de estas estadísticas dado que desde el 2001 hasta la fecha el APA ha insistido en reportarlas. Se recomienda a Cumming y Calin-Jageman (2017), así como al APA (2020) para ver detalles e implicaciones de estas estadísticas.

Tabla 7. Resultados de la prueba de independencia

Estadística	Residuales ajustados estandarizados	χ^2 Calculada	*p*
Promedio			
Reportado	2.7	7.29	0.0069
No reportado	-2.7	7.29	0.0069
SD			
Reportada	1.1	1.21	0.2712
No reportada	-1.1	1.21	0.2712
IC			
Reportado	-6.8	46.24	< 0.00001*
No reportado	6.8	46.24	< 0.00001*
Tamaño de efecto			
Reportado	-4.4	19.36	0.000011*
No reportado	4.4	19.36	0.000011*
Tipo de análisis			
Reportado	3.5	12.25	0.00047*
No reportado	-3.5	12.25	0.00047*
Estadística de la prueba (valor *t* o *F*)			
Reportada	1.9	3.61	0.0574
No reportada	-1.9	3.61	0.0574
Df			
Reportado	-0.5	.25	0.617
No reportado	0.5	.25	0.617
P			
Reportado	2.7	7.29	0.00693
No reportado	-2.7	7.29	0.00693

Nota: *resultados estadísticamente significativos bajo un $\alpha_{ajustado} = 0.0031$.

Fuente: Elaboración propia

Discusión

Trayendo de vuelta la pregunta de investigación, "¿cuál ha sido la calidad del reporte de algunas estadísticas de publicaciones educativas arbitradas y relacionadas a procesos experimentales?", y dados los criterios de poder, replicación, accesibilidad y el reporte de las estadísticas con diferencias estadísticamente significativas, la evaluación resultó ser considerada de mala calidad (siguiendo la antes citada escala de Kotz, 2006). Para el futuro, se pueden mejorar bastante las publicaciones con seguir estos criterios. Los objetivos se cumplieron al evaluar la calidad estadística de artículos arbitrados (n = 34) con diseños experimentales y enlistar una serie de recomendaciones para incrementar su calidad (APA, 2020; Cohen, 1988; Cumming y Calin-Jageman, 2017; Maxwell *et al*., 2018).

Al no detallar el poder estadístico, no se sabe la probabilidad de haber rechazado una H_0 falsa. Por ejemplo, con un poder de 80 %, al incrementar la probabilidad de que una H_0 sea falsa (i.e., un sinónimo es que la H_A sea cierta), la probabilidad de observar un VP incrementa ($p < \alpha$) y la probabilidad de un FP disminuye ($p \geq \alpha$) (Harms y Lakens, 2018). Con lo anterior, y un $p <$ alfa, se podría observar en las muestras si hubo un efecto estadístico de un tratamiento en un diseño experimental. Si así fuera, esto sería algo prometedor para replicarlo y observar si se repite el fenómeno (Feynman, 1974; Greenland *et al*., 2016; Harms y Lakens, 2018). Para hacer una réplica, es necesario contar con toda la información relevante, así que el dar acceso a bases de datos y análisis es una condición que se debe cumplir sí o sí, solo de esta forma es posible contribuir al avance del conocimiento y de la ciencia (Carey, 2011; Cumming, G. and Calin-Jageman, 2017). De otra manera, sin saber del poder ni seguir una replicación ni dar acceso a bases de datos y omitir estadísticas relevantes en publicaciones, los resultados de una investigación se pueden volver cuestionables (Cumming y Calin-Jageman, 2017). La APA (2020) ha recomendado declarar la estimación del poder y otros métodos usados para determinar la precisión de las estimaciones de los parámetros. El proceso de determinar el número de casos, participantes u observaciones que un estudio necesitaría para alcanzar un nivel de potencia deseado con un cierto tamaño de efecto y un cierto nivel de significancia para rechazar la hipótesis nula.

Los autores de la muestra en su mayoría incluyeron el promedio, SD, tipo de análisis, y p, y todos declararon el n. En contraste, la minoría calculó un df, IC y tamaño de efecto (tabla 3). Para observar si estas diferencias en las frecuencias se debieron al mero azar o había un efecto, se llevaron a cabo dos pruebas de ji al cuadrado: bondad de ajuste (tabla 6) y prueba de independencia (tabla 7). Los grados de libertad (df) tuvieron una diferencia estadísticamente significativa en la prueba de bondad de ajuste ($p = 0.001 < \alpha_{\text{ajustado}} = 0.0063$). Esta diferencia estadísticamente significativa fue evidencia de que el reporte de esta estadística fue descuidado y no fue probablemente al azar. Se recomienda su reporte porque sirve para identificar un valor F o t crítico para rechazar o no una hipótesis nula, así como para tener suficiente información para replicar un estudio.

Otras dos estadísticas que fueron preocupantes por no ser reportadas suficientemente fueron el IC y el tamaño del efecto. La razón fue que sus probabilidades calculadas fueron menores a los coeficientes alfa ajustados de la prueba de bondad (para el IC y tamaño del efecto su $p = 0.00001 < \alpha_{\text{ajustado}} = 0.0063$) y en la prueba de independencia (IC, $p < 0.00001$

y tamaño del efecto, p = 0.000011, que fueron menores al $\alpha_{ajustado}$ = 0.0031). Estas dos estadísticas también fueron descuidadas a pesar de haber sido identificadas como importantes desde el 2001 por el APA. Un IC ayuda a estimar el valor del parámetro de interés (e .g., un de 95 % significa que si se toman 100 muestras similares 95 de ellas contendrán el parámetro de la población y cinco no; y también se estima el error, que puede ser de algunos puntos, si se tiene un error de tres puntos un IC de 70 puntos en la muestra, se esperaría que el parámetro de la población estuviera entre 67 y 73 puntos). A menudo, los tamaños del efecto se interpretan como indicativos de la importancia práctica de un hallazgo de investigación, es decir, el grado en que el fenómeno está presente en la población o el grado en que la hipótesis nula es falsa (APA, 2020).

Aunque las y los autores no tocaron el tema del poder, este se estimó con alguna de la información de los artículos y asumiendo otras para el presente estudio. La calculadora de G*Power necesita para una prueba *a priori* (Faul *et al*., 2009) los siguientes datos:

- El coeficiente de *d* (el *d* de Cohen; ver a Cumming y Calin-Jageman, 2017;
- El coeficiente de α;
- Poder deseado (por lo menos 80%) y
- El tamaño de los grupos (*n*);

Por otro lado, G*Power necesita para la prueba *a posteriori*: *d*, α y *n* (Faul *et al*., 2009).

Una vez estimado el poder de los 34 artículos de la muestra, se obtuvo 41.2 % por debajo y 58.8 % por encima de 80 %, pero estos números se tienen que tomar con precaución porque solo 35.29 % estimó el tamaño del efecto, así que se *asumió* un tamaño efecto mediano para estas estimaciones de poder. Similar a Ioannidis (2005), la falta de poder en las estimaciones de estos artículos de investigación educativa hace *cuestionable* sus conclusiones para identificar VP (e. g., estudiantes que verdaderamente se hayan beneficiado de un tratamiento de tutorías) y FP (e. g., estudiantes que falsamente se hayan beneficiado de un tratamiento de tutorías). La realización de un análisis de poder implica decidir *a priori: a)* cuál será el tamaño del efecto esperado o qué tamaño del efecto será significativo en función de lo que se encontró en investigaciones anteriores o importancia teórica o práctica, *b)* a qué nivel de *p* se rechazará la hipótesis nula y *c)* qué probabilidad será suficiente para rechazar la hipótesis nula si realmente existe una relación en la población (el poder del estudio) (APA, 2020).

Importancia y significancia de los resultados y su aplicación e impacto en los campos del conocimiento

Es por esto por lo que la significancia de los resultados y su aplicación e impacto en los campos del conocimiento resulta verdaderamente importante, sobre todo en esta época de la COVID-19, que ha tocado casi cada aspecto de la vida moderna, pues, así como se espera que la medicina encuentre la solución al problema de la pandemia, las expectativas de calidad y funcionamiento en la educación deber ser las mismas y ser sometidas a evaluaciones del más alto rigor científico. Desde hace tiempo, Ioannidis (2005) afirmó en la investigación médica que la mayoría de los resultados son falsos, principalmente por la falta de poder.

Similarmente, los artículos de la muestra de la presente investigación carecieron de incluir varias estadísticas fundamentales, lo que hace que las inferencias hechas sean sumamente cuestionables, por decir lo menos.

Un poder de 80 % quiere decir que cuatro de cinco personas fueron identificadas correctamente como verdaderos positivos cuando $p < \alpha$. No basta con que se haya encontrado en un estudio un $p < \alpha$, habría que repetirlo en muchas ocasiones y publicarlo, aunque no fuera significativo estadísticamente. Luego se podría hacer un metaestudio para ver si se puede observar un patrón en el fenómeno en cuestión (Cumming y Calin-Jageman, 2017).

Generalización

Ya que se hicieron dos pruebas de inferencia estadística (de bondad y de independencia), los resultados de la presente investigación se podrían extender a las revistas (20) de la Redib bajo las siguientes categorías de búsqueda: "Ciencias sociales y humanidades" y "Educación e investigación educativa", tanto en inglés como en español. Otra distinción de los artículos de la muestra fueron las palabras clave *experimento* y *experiment*, así como el haber llevado un diseño experimental y un análisis paramétrico.

Limitaciones y recomendaciones

Entre las principales limitaciones se encontraron no haber evaluado el método de muestreo usado en los artículos; no haber examinado los supuestos de los análisis paramétricos (e. g., distribución normal, valores perdidos, linealidad, homogeneidad de las varianzas y observaciones atípicas, entre otras) (Maxwell *et al.*, 2018), y no haber cubierto los valores *p* en detalle con respecto a la inflación del error tipo I, entre muchos otros más. Por ello, se recomienda cubrir estos temas, replicar el presente estudio y ahondar en los tópicos del tamaño del efecto (Cumming y Calin-Jageman, 2017).

Recomendaciones de políticas de las revistas

Antes que nada, se insta a las revistas que publican artículos con diseños experimentales a que colaboren con el Center for Open Science. Esto les puede ayudar a publicar artículos con resultados rigurosos en lugar de resultados cuestionables, como los que abundan en la muestra del presente estudio. Otra ayuda fundamental para manuscritos como los anteriores es el uso del manual de la APA (2020) y la consulta de autores como Cohen (1988), Cumming y Calin-Jageman (2017) y Maxwell *et al*., (2018), entre muchas otras publicaciones que se mencionan en el presente estudio. Estos textos no solo deberían ser usados por los autores, sino también por las y los árbitros de las revistas para poder fundamentar de una manera más sólida y válida las inferencias que se hacen de los resultados.

Conclusiones

Los autores y autoras de la muestra de 34 artículos *no* cumplieron con los criterios (estipulados por el presente estudio) de poder, replicación, accesibilidad y solo parcialmente con el reporte de algunas estadísticas. En unas cuantas palabras, la conclusión del presente estudio fue que la calidad se considera *mala* por parte de esta muestra de publicaciones. Estos cuatro criterios de evaluación se tomaron de la literatura antes mencionada y ayudan a hacer más robustas las inferencias que se hagan de un estudio de corte experimental. Según algunos y algunas investigadoras mencionados en este estudio, estos criterios forman parte de las *mejores prácticas* para hacer investigación empírica tanto en las ciencias sociales como en las demás ciencias. Para robustecer las inferencias y avances de la ciencia misma, se recomiendan los siguientes tres puntos basados en los criterios de evaluación del presente estudio: *1)* calcular el poder estadístico para poder rechazar una hipótesis nula falsa y poder minimizar el número de falsos negativos, *2)* no solo buscar hacer siempre nuevos estudios, sino *replicar* los ya existentes para observar si se forma algún patrón a lo largo del tiempo con algún efecto: i. e., tener evidencia de algún efecto que se repita y *3)* no guardarse los datos y los análisis, sino dar accesibilidad a los interesados para replicar y encontrar potenciales errores y reportar las estadísticas: promedio, desviación estándar, intervalo de confianza, tamaño del efecto, tipo de análisis, estadística de prueba *t* o *F* (entre otras), grados de libertad y probabilidad calculada. Asimismo, habría que consultar el APA (2020) para ver qué otras estadísticas son pertinentes de acuerdo con el tipo de estudio.

Futuras líneas de investigación

Es necesario seguir analizando la literatura de investigación educativa en lo que se refiere a diseños experimentales para ver la calidad con la que cuenta. Además, cabe la posibilidad de que se pudieran evaluar los diseños *no experimentales* en cuestión del nivel de poder, ver si son parte de una serie de réplicas y observar el acceso a sus datos y análisis, así como el reporte de sus estadísticas. Un primer posible paso sería examinar la representatividad de la muestra de estudio ante una población para una posible generalización de lo encontrado en los análisis. Es decir, si la muestra fue tomada al azar de una población y tiene el tamaño mínimo que se marca por alguna fórmula con un nivel de confianza (e. g., 95 % o 99 %) y con cierto margen de error (intervalo de confianza). Cuando una muestra es de conveniencia, se puede hacer el argumento de que tiene estadísticas de interés similares a los parámetros de una población. Lo anterior se podría establecer con una prueba de significancia estadística para comparar grupos y correlacionar variables, así como un análisis de tamaño de efecto. De este modo, se tendría un respaldo más sólido para la generalización de los resultados de un estudio.

Otros aspectos que se podrían revisar de la literatura de investigación educativa son las *propiedades psicométricas* de los instrumentos de recolección de datos (exámenes y encuestas; i. e., validez y confiabilidad de los puntajes). Es igualmente necesario cubrir las propiedades psicométricas para ver qué tanta coherencia hay entre las mediciones (confiabilidad), así como si se está midiendo lo que se desea de medir (validez). Si los puntajes de los instrumentos *no* exhiben niveles de confiabilidad y validez apropiados

(mínimos), no tiene caso continuar con los análisis estadísticos y tamaños de efecto. Por último, se podría replicar el presente estudio para evaluar la calidad de los artículos en otros portales y revistas tanto con diseños experimentales como con no experimentales.

Referencias

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine, 18*(3), 91-93. Retrieved from https://doi.org/10.1016/j.tjem.2018.08.001.

American Psychological Association [APA]. (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, United States: American Psychological Association.

American Psychological Association [APA]. (2020). *Publication Manual of the American Psychological Association: The Official Guide to APA Style* (7th ed.). Washington, United States: American Psychological Association.

Armstrong, R. (2014). When to use the Bonferroni correction. *Ophthalmic & Physiological Optics, 34*(5), 502-508. Retrieved from https://onlinelibrary.wiley.com/doi/10.1111/opo.12131.

Beasley, T. M. and Schumacker, R.E. (1995). Multiple Regression Approach to Analyzing Contingency Tables: Post Hoc and Planned Comparison Procedures. *The Journal of Experimental Education, 64*(1), 79-93. Retrieved from https://doi.org/10.1080/00220973.1995.9943797.

Berenson, M, Levine, D., Szabat, K. and Stephan, D. (2019). *Basic Business Statistics: Concepts and Applications* (14th ed.). New York, United States: Pearson.

Carey, S. (2011). *A Beginner's Guide to Scientific Method* (4th ed.). New York, United States: Wadsworth Cengage Learning.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York, United States: Psychology Press.

Cumming, G. and Calin-Jageman, R. (2017). *Introduction to the New Statistics: Estimation, Open Science, and Beyond*. New York, United States: Routledge.

Ellenberg, J. (2014). *How not to be wrong: The power of mathematical thinking*. New York, United States of America: Penguin Press.

Faul, F., Erdfelder, E., Buchner, A. and Lang, A. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. Retrieved from http://dx.doi.org/10.3758/BRM.41.4.1149.

Feynman, R. (1974). Cargo Cult Science. *Engineering & Science, 37*(7), 10-13.

Fisher, R. (1949). *The design of Experiments.* New York, United States of America: Hafner.

Gall, M., Gall, J. and Borg, W. (2007). *Educational Research: An introduction* (8th ed.). New York, United States: Pearson.

Greenland, S., Senn, S., Rothman, K., Carlin, J., Pole, C., Goodman, S. and Altman, D. (2016). Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31*, 337-350. Retrieved from http://dx.doi.org/10.1007/s10654-016-0149-3.

Greenwood, P. and Nikulin, M. (1996). *A Guide to Chi-Squared Testing*. New York, United States: John Wiley & Sons.

Hancock, G., Stapleton, L. and Mueller, R. (2019). *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (2nd ed.). New York, United States: Routledge.

Harms, C. and Lakens, D. (2018). Making 'Null Effects' Informative: Statistical Techniques and Inferential Frameworks. *Journal of Clinical and Translational Research, 2*, 382-393. Retrieved from https://doi.org/10.17605/OSF.IO/WPTJU.

Hinkle, D. E., Wiersma, W. and Jurs, S. G. (2003). *Applied Statistics for the Behavioral Sciences*. United States: Houghton Mifflin Harcourt.

Ioannidis, J. (2005). Why Most Published Research Findings are False. *PLoS Medicine*, *2*(8), 696-701. Retrieved from https://doi.org/10.1371/journal.pmed.0020124.

Kohler, H. (2020). *Hypothesis Testing: The Chi-Square Technique (Statistics: A Universal Guide to the Unknown Book 14).* Amherst, United States: Heinz Kohler.

Kotz, S. (2006). *Encyclopedia of Statistical Sciences* (2nd ed.). New Jersey, United States: Wiley-Interscience.

Lakens, D. (n. d.). Improving your statistical inferences. (MOOC). Coursera. Retrieved from https://www.coursera.org/learn/statistical-inferences/lecture/erVLS/type-1-and-type-2-errors.

Maxwell, S., Delaney, H. and Kelley, K. (2018). *Designing Experiments and Analyzing Data* (3rd ed.). New York, United States: Routledge.

Meehl, P. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychology Inquiry, 1*(2), 108-141. Retrieved from http://dx.doi.org/10.1207/s15327965pli0102_1.

Nicol, A. and Pexman, P. (2010). *Presenting your Findings: A Practical Guide for Creating Tables* (6th ed.). Washington, United States: American Psychological Association.

Ponce, H. (2019). *Conceptos básicos de estadísticas inferenciales aplicadas a la investigación educativa*. Ciudad Juárez, México: Universidad Autónoma de Ciudad Juárez.

Russo, R. (2021). *Statistics for the Behavioral Sciences: An Introduction to Frequentist and Bayesian Approaches* (2nd ed.). London, England: Routledge.

Salkind, N. (2007). *Encyclopedia of Measurement and Statistics, Volume 1*. New York, United States: Sage.

Sakai, T. (2018). *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes and Statistical Power*. Singapore: Springer.

Sarkar, S. and Pfeifer, J. (2006). *The Philosophy of Science: An Encyclopedia*. New York, United States: Routledge.

Singh, P. and Khan, B. (2019). *Writing Quality Research Papers: Brief Guidelines to Enhance the Quality of Research Paper/Manuscript*. Mumbai, India: BPB Publications.

VandenBos, G. (2015). *APA Dictionary of Psychology* (2nd ed.). Washington, United States: American Psychological Association.

Rol de Contribución	Autor (es)
Conceptualización	Héctor Francisco Ponce Renova (principal) Diana Irasema Cervantes Arreola (igual) Beatriz Anguiano Escobar (igual)
Metodología	Héctor Francisco Ponce Renova (principal) Diana Irasema Cervantes Arreola (igual) Beatriz Anguiano Escobar (igual)
Software	NO APLICA
Validación	Héctor Francisco Ponce Renova (principal) Diana Irasema Cervantes Arreola (igual) Beatriz Anguiano Escobar (igual)
Análisis Formal	Héctor Francisco Ponce Renova (principal) Diana Irasema Cervantes Arreola (igual) Beatriz Anguiano Escobar (igual)
Investigación	Héctor Francisco Ponce Renova (principal) Diana Irasema Cervantes Arreola (igual) Beatriz Anguiano Escobar (igual)
Recursos	Héctor Francisco Ponce Renova (principal) Diana Irasema Cervantes Arreola (igual) Beatriz Anguiano Escobar (igual)
Curación de datos	Héctor Francisco Ponce Renova (principal) Diana Irasema Cervantes Arreola (igual) Beatriz Anguiano Escobar (igual)
Escritura - Preparación del borrador original	Héctor Francisco Ponce Renova (principal) Diana Irasema Cervantes Arreola (igual) Beatriz Anguiano Escobar (igual)
Escritura - Revisión y edición	Héctor Francisco Ponce Renova (principal) Diana Irasema Cervantes Arreola (igual) Beatriz Anguiano Escobar (igual)
Visualización	Héctor Francisco Ponce Renova (principal) Diana Irasema Cervantes Arreola (igual) Beatriz Anguiano Escobar (igual)
Supervisión	Héctor Francisco Ponce Renova (principal) Diana Irasema Cervantes Arreola (igual) Beatriz Anguiano Escobar (igual)
Administración de Proyectos	Héctor Francisco Ponce Renova (principal) Diana Irasema Cervantes Arreola (igual) Beatriz Anguiano Escobar (igual)
Adquisición de fondos	NO APLICA