Ride 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.

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

Introduction

This work is an evaluation of a sample of articles with experimental designs that were published in peer-reviewed journals. Making a reference to COVID-19 and experimental designs, it has been essential to differentiate infected people, detect their contacts and provide isolation, as well as apply effective treatments and vaccination. To be successful, experimental and statistical processes (among many other resources) are required to differentiate the effectiveness of a screening test and an effect (i. E., Vaccines and treatments). Now, the scientific method is not only used to solve health problems, but also to address educational problems (eg, effect of tutorials, distance learning, didactic strategies, among others). In this regard, Maxwell, Delaney and Kelley (2018) explain that "experimental design and data analysis methods derive their value from the contributions they make to the general activity of science" (p. 3). In a similar way to medicine, this text analyzes some parts of the quality of statistical analyzes used in experimental processes in educational research. Feynman (1974), during a speech to a generation of recent graduate students, stressed the importance of revealing all pertinent information and being careful when doing science:

> The first principle is that one should not fool oneself — and one is the easiest person to fool. So you have to be very careful about it. After not having fooled yourself, it is easy not to fool the other scientists. One just has to be honest in the conventional way after all (p. 12).

As will be seen in detail later, if a study does not have sufficient statistical power (Ponce 2019), a researcher, by ignoring an undetected effect due to lack of power, could infer that the results are not statistically significant and arrive to self-deceive.

On the other hand, the present study deals with statistical inferences and not theoretical inferences (Meehl, 1990). Statistical inferences start from a sample of participants, observations or objects to generalize a series of results to the corresponding population and include the following elements: power, hypothesis testing (eg, test to F), calculation of a confidence interval and estimation of an effect size (Cohen's d) (Cumming and Calin-Jageman, 2017).

It is about offering a set of resources for educational researchers in the area of experimental designs and their corresponding statistics. The research question of the present study was: what has been the quality of the reporting of some statistics of refereed educational publications related to experimental processes? One of the objectives was derived from the question: to evaluate the quality of some statistical processes of the experimental educational publications published. The second objective was to give recommendations to increase its quality. This quality had the following criteria:

1) Have discussed, calculated and obtained enough power to reject a false null hypothesis.
2) Observe if the study in question is part of a series of replications.
3) See if the databases were given access for potential replicas.
4) Report statistics suggested by the American Psychological Association [APA] (2001, 2020) from 2001 to the present day.

Theoretical framework

In this section, some aspects related to the above-mentioned criteria will be detailed. Regarding the first of them, the weight of statistical power, the APA (2020) explains the following:

> When applying inferential statistics, take seriously the considerations related to the statistical power associated with testing for hypotheses. Such considerations relate to the probability of correctly rejecting the hypotheses tested given a given alpha level, an effect size, and a sample size. In this regard, evidence must be provided that the study has sufficient power to detect effects of substantive interest. (p. 86).

Regarding the second criterion, Feynman (1974) also suggests, instead of insisting on something new each time, repeating experiments to see if the cause and effect of a phenomenon were found. In addition, it would be necessary to see if the phenomenon repeats itself before believing that one has something: i. e., p-values were conceptualized in the long term and not for a single occasion (Greenland et al., 2016; Harms and Lakens, 2018).

Regarding the third criterion, it would be necessary to see if the databases were given access. For example, the Center for Open Science was created in 2013 to promote the openness, integrity, and reproducibility of scientific research (Cumming and Calin-Jageman, 2017). It is a non-profit organization sponsored by government and private funds with approximately 205 subscribed magazines. In the first stage, the process is free and involves sending a protocol with an introduction, methods and the results of some piloting. The manuscript is then evaluated by editors and reviewers, who can provide feedback. Under the condition of following the protocol in the experiment, the journal in question guarantees the publication of the manuscript. In the second stage, the manuscript should include the introduction, the methods, the results of the new analyzes and the discussion. Authors may be asked or required to share their data sets in a free public file to access and are encouraged to share the code for their statistical analyzes. After this process, the finished article is published. Finally, a log report is published and will appear to reassure readers that the main hypotheses and analyzes are free from questionable research practices. Soler (2016) complements that "said document must contain sufficient information that allows other researchers on the subject to understand the advances described, evaluate the results and understand the scope of the conclusions" (p. 4).

The fourth criterion was to reiterate some of the demands and suggestions of agencies (APA) and authors such as Cohen (1988), Cumming and Calin-Jageman (2017) and Maxwell et al. (2018) to document some statistics in experimental designs, as well as explain the procedures behind them. According to the APA (2020):

> When describing inferential statistics (e.g., t or F tests associated with effect size and confidence intervals), enough information is included to allow the reader a complete understanding of the analyzes that were carried out. The data provided, preferably in the text, but possibly in supplementary materials depending on the magnitude of the data set, should allow the reader to confirm the basic reports of the analyzes (eg, means, SD, sample size, correlations) and should empower the interested reader to construct estimates of effect

> sizes and confidence intervals beyond those provided in the manuscript per se (p. 181).

Ioannidis (2005) shows that the probability of a scientific statement being true / true depends on the following variables:

- Power.
- Biases (manipulation of analysis or in the results section, and selection or distortion of the results is a typical form of bias).
- The number of studies answering the same research question (a small number makes the results less likely to be true).
- Ratio of true relationships to no relationships in investigations (the higher the ratio of true relationships to no relationships [eg, 2: 1] the greater the probability of finding true relationships).

Specifically, Ioannidis (2005) states that "published research results are sometimes refuted by subsequent evidence, ensuring confusion and disappointment" (p. 696). Also, that most of the findings are false and this is demonstrable. It should be noted that these statements were issued in the context of medicine and it remains to be seen if they are equally applicable in the field of educational research.

Academic justification: gap in the literature

During the spring of 2020, the search for the question of the present investigation produced zero articles in Google Scholar. In contrast, in this same search engine, four publications were found using the keywords of the question: statistics, educational research and experimental processes. Three of them were theses and the other a compilation of scientific research in engineering and education. None of them covered the question or the objectives of this one. Given the results of Google Scholar, a gap in the literature is assumed.

Here we try to fill that gap with the results and analysis of the sample. The above justifies this research: filling the gaps in knowledge contributes to better describe, explain and predict reality (Gall, Gall and Borg, 2007). Indeed, science does not aim at anything other than "to increase our understanding of why things happen the way they do" (Carey, 2011, p. 2). To do science, you can use the APA guide (2020). Another alternative to describe statistics related to experimental designs was given by Nicol and Pexman (2010).

Practical justification

The results of this study can help to follow certain principles to detail the results of experimental investigations (i. E., Power, replica, access to your databases and statistical reports). This can contribute to making inferences better grounded in statistical theory and to creating accessible, transparent processes with the necessary information so that other researchers can make replicas.

Conceptual and operational definitions

In this work, the frequency or classical statistic is used because of its great use and because it corresponds to the sample (consult Russo [2021] for more information on the frequency theory). The test of statistical significance of the null hypothesis uses the calculated probability $p < \alpha$ (decision: reject the null, H0) and $p \geq \alpha$ (decision: do not reject H0). The p corresponds to the data and is the probability of finding a certain test statistic (eg, a t or F value) or a more extreme one, when H0 is true. According to Salkind (2007), this test can be useful to explore the following questions:

a) The uncertainty inherent in the empirical data.

b) The nature of statistical inferences.

c) The test statistic (eg, t or F value), which represents a result of an analysis.

d) The nature of a decision to reject an H0 in terms of a random effect.

Quality

In everyday life, a user compares products and decides their quality: he qualifies what he is comparing as bad, good and excellent (Kotz, 2006; Singh and Khan, 2019). The way to make this concept operational is to subject it to the four criteria described above.

Other definitions

Gall *et al.* (2007) They stated that the experiment is the most powerful quantitative research method for establishing cause and effect relationships between two or more variables. Due to this ability to establish the cause and effect relationship, this design was selected for the present study. This experimental design corresponds to the analyzes proposed by the APA and covered by Cohen (1988), Cumming and Calin-Jageman (2017) and Maxwell et al. (2018), among others. Said cause and effect relationship in an experimental design (DE) can be questionable due to internal and external threats (see Gall et al. [2007] to delve into these threats).

The characteristics of the sample selected here do not fully match this definition of experiment. One of the reasons was that in most articles the respective population samples were not randomly selected, as well as the control group (received no treatment) and the treatment / experimental group (received the treatment). Therefore, another more appropriate definition was used:

> [The] experimental design is a plan of the procedures to be followed in scientific experimentation to reach valid conclusions, with considerations of such factors as selection of participants, manipulation of variables, data collection and analysis, and minimization of external influences. (VandenBos, 2015, p. 397).

An experimental design involves a cause and an effect: an event or state that is brought about as a result of another (VandenBos, 2015). In other words, the cause or independent variable is measured on a nominal scale (eg, sex: male and female) and is manipulated: the sample is divided into a control group (without treatment or with a placebo; eg, tutorials) and a treatment group (eg, with tutorials). The effect can be measured with the difference between independent groups and within dependent groups (pre and post-test). To observe the possible effect between these variables, a statistical test is used to test the H0 (eg, independent or dependent samples t-test and one-factor analysis of variance). The H0 establishes that there is no difference between the population averages, as well as between the samples (Russo, 2021; Salkind, 2007).

Whether for analysis of independent or dependent groups, the effect becomes operational by obtaining its magnitude: “Effect size, which is one or more measurements of the magnitude or meaning of the relationship between two variables. Next, the effect sizes are interpreted as indicative of the practical significance of a research result ”(VandenBos, 2015, p. 352). For the present investigation, the main effect was taken into account. VandenBos (2015) defined it as follows: “The consistent total effect of a single independent variable on a dependent variable on all other independent variables in an experimental design. This is different from, but can be obscured by, an interaction effect between variables ”(p. 617).

In his seminal book, Cohen (1988) gave various sizes to classify the effects: small, medium, and large; but he cautioned that these sizes should only be used when there is no context to be interpreted. To delve into other effect sizes, Hancock, Stapleton, and Mueller (2019) and Sakai (2018) are recommended for laboratory experiments. It is beyond the objectives of this present manuscript to assess the effect size of the sample.

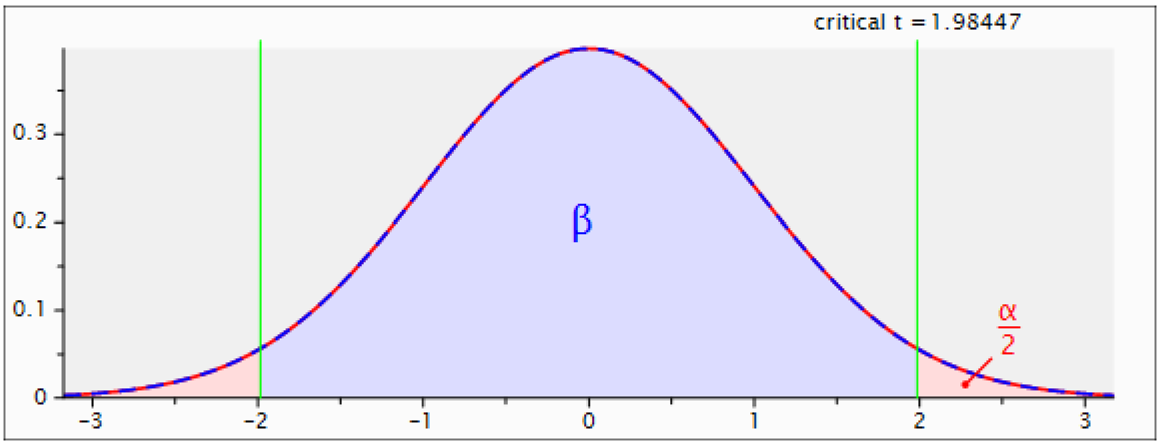
Statistical significance

Figures 1-6 contain a series of examples to explain the concepts of statistical significance, null hypothesis (H0) and alternative (HA), α (level of significance; type I error [TSI]), beta (β; type II error [ETII]), power, samples, populations and sampling error. Statistical significance is “the degree to which a research result cannot be reasonably attributed to the intervention of chance or other random factors” (VandenBos, 2015, p. 1026). With the statistical significance test, H0 is tested: the phenomenon does not exist (Fisher, 1949).

Delving into this, Ellenberg (2015) explained: “A statistically significant result gives us a clue, suggesting a promised land to focus investigative energy” (p. 156). In more detail, this author mentioned that a test of statistical significance functions as if it were a detective, not a judge per se. In other words, a p <α is the beginning of some promising result with a series of replications (Cumming and Calin-Jageman, 2017), but not an end in itself. If H0 is rejected, it is concluded that the phenomenon exists (i. E., Observable event, according to VandenBos [2015]).

So, the p-value indicates how surprising the results of the analyzes are under the assumption that there is no effect (figure 1; the area shaded in blue is the non-rejection area of the null: 1 - α: commonly it is 1 - 0.05 = 0.95). This means that if you compare the averages of a math test in two groups, they will probably not be the same. If the difference between these means is not very large, p ≥ α, these results are not surprising and the difference was due to chance. On the other hand, if p <α, the results would be surprising, under the assumption that there is no difference between the populations: i. e., there is the possibility of some effect. In summary, when the null is true (Figure 1; there is no true effect: the distributions of both groups perfectly overlap), the probability of finding statistically significant results is equal to α.

Figura 1. Dos distribuciones normales perfectamente superpuestas

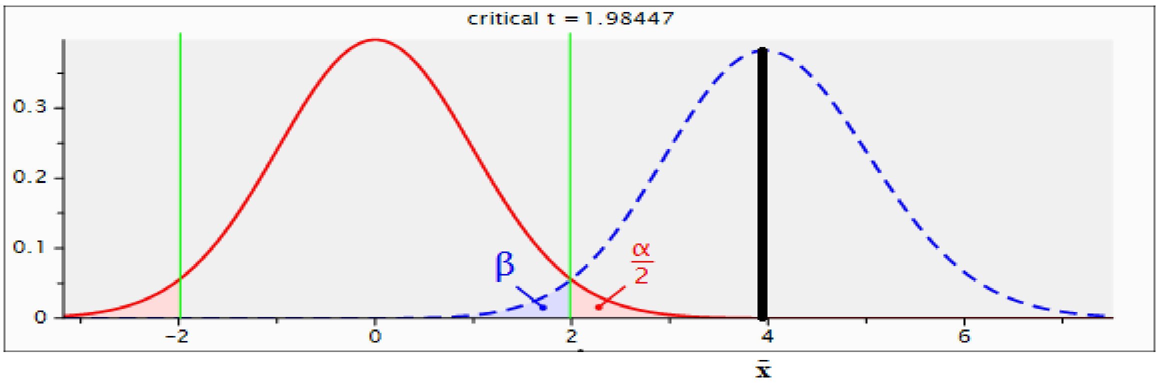


Nota: Modelo de rechazo de la H_0, así como una H_0 cierta. Se aplicó una prueba *t* de grupos independientes dado: poder = 0.05; α = 0.05/2; β = 0.95; *t* crítico = 1.98, *df* = 98 y *p* > α.

Fuente: Elaboración Propia

On the other hand, if the null is false (figure 2; there is a true effect) the probability of finding statistically significant results is equal to the power (eg, 0.80 or 80% recommended as a minimum by Cohen [1988] and Cumming and Calin-Jageman [2017]). In fact, Cohen (1988) simply defined it as follows: "The power of a statistical test is the probability that this test will give significant results" (p. 1). Simply put, power is the probability of obtaining significant results when the null is false.

Figura 2. Dos distribuciones normales separadas



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

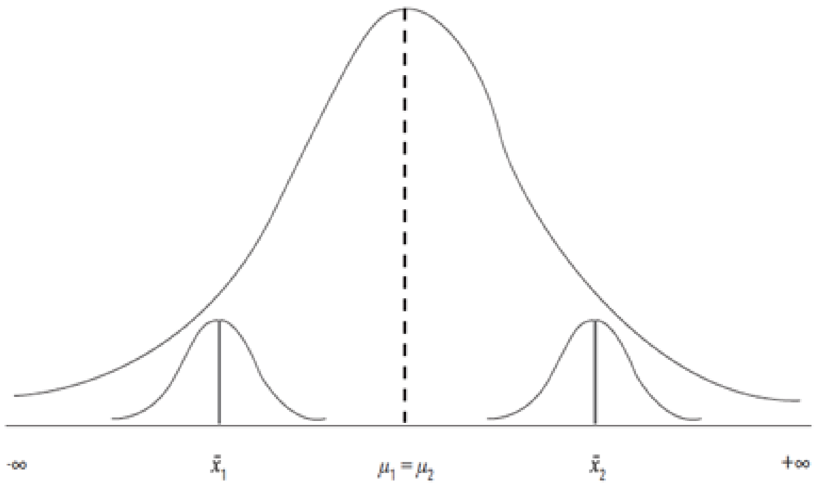
Fuente: Elaboración propia

Assuming that a researcher carries out an experimental design (Study I), she takes a random sample (n = 100) from the population (it must be remembered that parameters are obtained from the populations: eg, μ = mean of the population, σ = standard deviation of the population and δ = Cohen's d of the populations) and obtain statistics from the samples (x̄ = mean of the sample, SD = standard deviation of the sample and d = Cohen's d of the samples). She randomly divides the sample in two to have a control group (50) and a treatment group

(50); implements an independent samples t test with $\alpha = 0.05$ divided into two tails $\alpha / 2$: $\alpha =$ also represents type I error: probability of rejecting an H0 when it is true. She assumes that both samples come from different population distributions (figure 1), but they are identical because their parameters are the same (H0 is true): i. e., $\mu 1 = \mu 2$ and $\sigma 1 = \sigma 2$, and therefore $\delta = (\mu 1 - \mu 2) / \sigma = 0$. The area of the two overlapping distributions is equal to one or 100%. On the other hand, there is the beta value (probability of not rejecting an H0 when it is false: 1 - power): in this case it turns out that $\beta = 1 - 0.05 = 0.95$ (figure 1), and the power is $1 - \beta$ ($1 - 95 = 0.05$), as well as $\alpha = 0.05$.

Something that can happen due to sampling error (difference error between the population parameters and the sample statistics) is shown in figure 3: H0 is true, but the two samples ($\bar{x}\ 1 \neq \bar{x}\ 2$, $d \neq 0$; $p < \alpha$). Based on the statistics of these samples, H0 is wrongly rejected (ie, type I error).

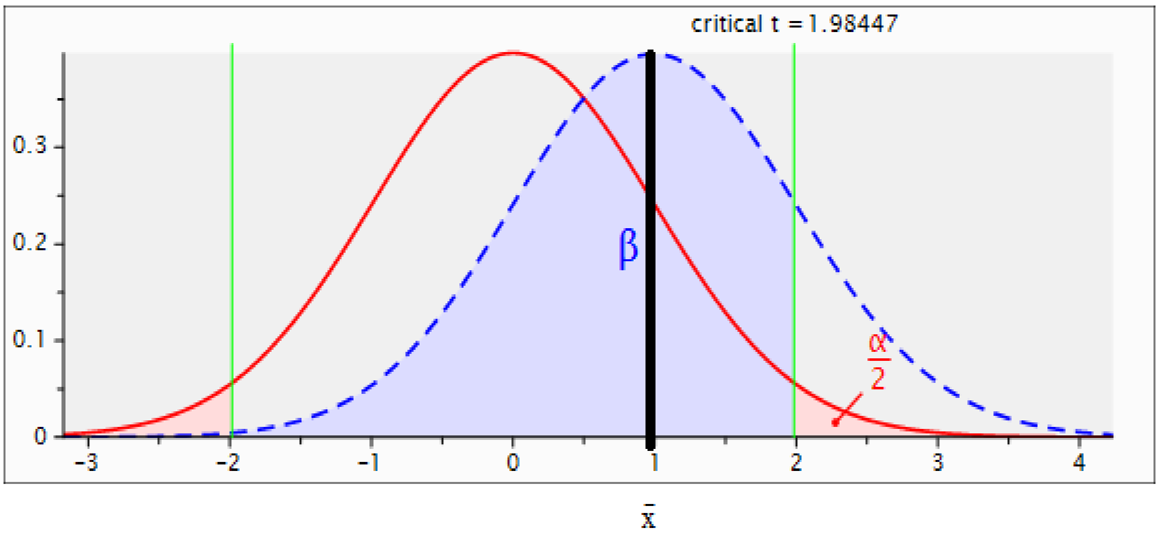
Figura 3. H_0 es cierta con error tipo I



Fuente: Ponce (2019)

Returning to the example of the researcher, she applies a treatment to one group and nothing to the other, and starts from a certain H0. As a result, figure 4 is obtained, where it is observed how the two samples represented by the two normal distributions (the solid curve represents the distribution of the control group sample and the dotted curve to that of the treatment group) do not overlap ($\bar{x}$ control $\neq$ $\bar{x}$ treatment): the treatment caused the distributions to separate with $d = 0.20$ and the power ≈ 0.17 ($\beta \approx 1 - .17 \approx .83$), but the $\bar{x}$ treatment did not fall in the rejection area of the null (shaded area of the tails: $p \geq \alpha$). In this case, it cannot be known whether it was due to the lack of a true treatment effect or not having a large enough sample with the power to detect it. However, to draw conclusions about the effectiveness of the practical effect of the treatment, one would have to look at the literature to draw any conclusions. (Cohen, 1988; Cumming y Calin-Jageman, 2017).

Figura 4. Dos distribuciones separadas por 0.20 de SD

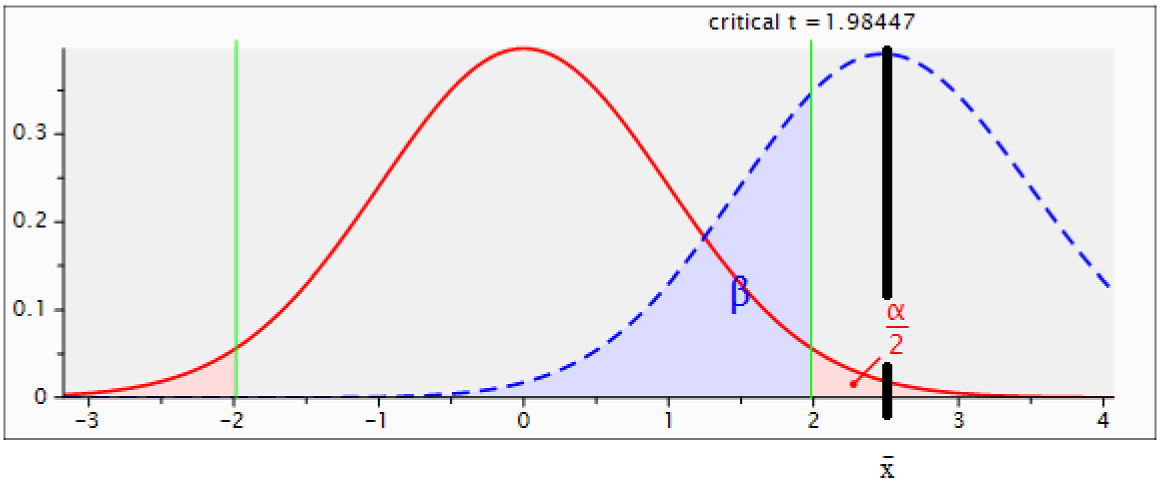


Nota: d cambió a 0.20; poder ≈ 0.17; $\alpha = 0.05/2$; $\beta \approx 0.83$; t crítico = 1.98, $df = 98$. La curva punteada representa, junto con la curva sólida, la hipótesis alternativa (H_A: $\mu_1 \neq \mu_2$).

Fuente: Elaboración propia

Another researcher replicates Study I (figure 5); find a d = 0.50 with a $p < \alpha$, $\beta \approx 0.30$, and the average of the treatment group falls in the area of rejection of the null. However, the power (≈ 0.70) it does not reach 0.80, so it would be recommended to increase it, which would increase the sample size.

Figura 5. Dos distribuciones separadas por 0.50 de SD

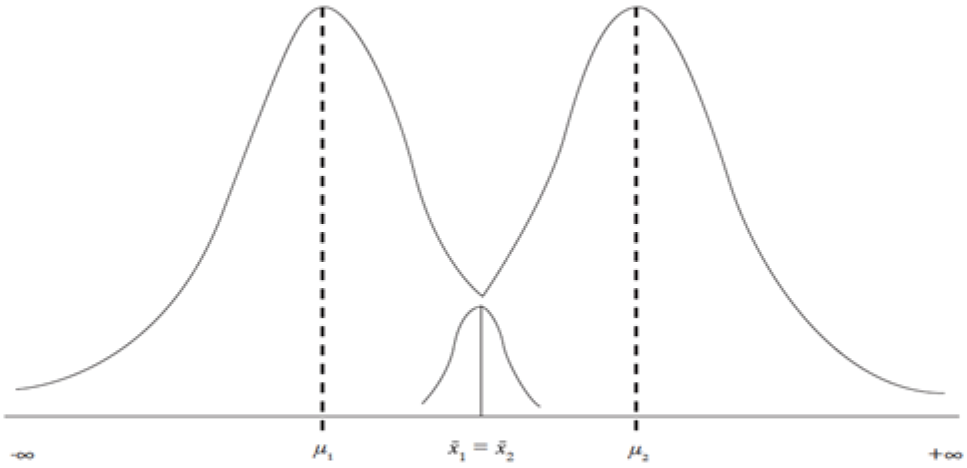


Nota: $d = 0.50$; Poder ≈ .70; $\alpha = .05/2$; $\beta \approx .30$; t crítico = 1.98, df = 98y $p <$ alfa.

Fuente: Elaboración Propia

A third researcher carries out another replication of Study I. His treatment has a d = 0.80 with $p < \alpha$, $\beta \approx 0.03$, and a power ≈ 0.97, so it had enough power: this would be a surprising result to investigate further (figure 2). Finally, figure 6 proposes that the samples ($\bar{x}_1 = \bar{x}_2$) when the H_0 it was false ($\mu_1 \neq \mu_2$), so a type II mistake is made by not rejecting a H_0 fake.

Figura 6. H_0 falsa con error tipo II



Fuente: Ponce (2019)

Implications of error I and II

One implication of the Type I error is obtaining a false positive [FP] (eg, identifying a student with outstanding abilities when in fact she does not). Likewise, one implication of type II error is finding a false negative [FN] (eg, a student does not qualify for special education when he should have been seen). This probability of obtaining a FP can be calculated by estimating the probability of a H_0 to be true or false, given a α. If one H_0 has a 50% probability of being true and a α = 0.05 or 5 %, the probability of a FP is 50% × 5% = 2.5% (i. e., 0.025). Table 1 shows the other probabilities of obtaining FP, as well as a true positive [VP] (eg, a student with outstanding abilities who is identified as such) and a true negative [VN] (a student without outstanding abilities who is identified as which does not have them).

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.)

Relationship between power and other variables

Power can be increased by increasing the sample size (Cohen, 1988; Cumming and Calin-Jageman, 2017). Figures 7-11 show the relationships between the following variables: power and FP, FN, VN and VP when the H_0 has a certain probability of being true or false. The following five examples show these relationships between these variables.

1) Power has an inversely proportional relationship to the beta coefficient (β): i. e., when the power increases 1%, the beta (FN) decreases 1% (figure 7). Power has no effect on α, but increasing α does increase power (ver a Cohen, 1988; Ponce, 2019).

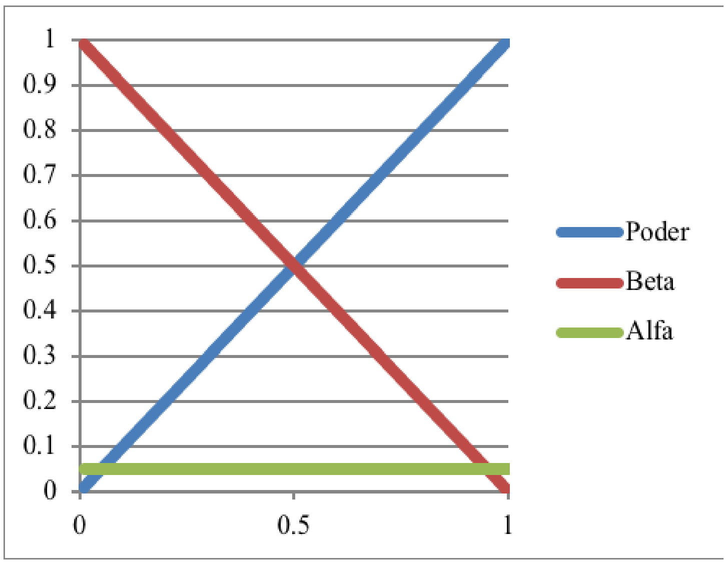
2) Given the H_0 with 50% being true, even though the power increases (figure 8), a given FP $p < \alpha$ remains fixed as does the NP given a $p \geq \alpha$ (see table 1 for mathematical calculations of probability).

3) When the H_0 has 50% of being false (figure 9), when increasing the power, the probability of a VP increases (when $p < \alpha$) and the probability of an FN decreases (when $p \geq \alpha$).

4) When the probability of the H_0 being true increases (figure 10), having the power constant at 80%, the probability of a FP increases, but its slope is less than the other probability that also increases, a VN. The decreasing probabilities are those of VP and FN (the latter has a smaller slope).

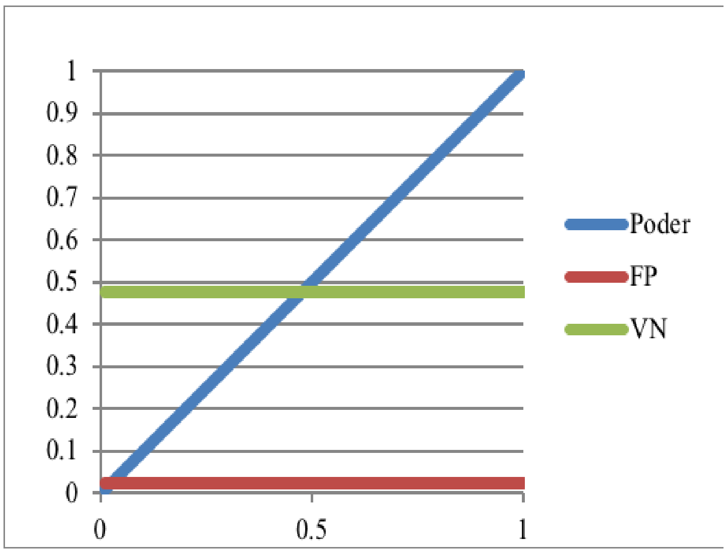
5) When the probability that an H0 is false increases (figure 11), keeping the power at 80%, the probability of a FP and a VN decrease (the latter has a steeper slope). On the other hand, the probability of a VP (greater slope) and a FN increase.

Figura 7. Relación entre el poder y el beta



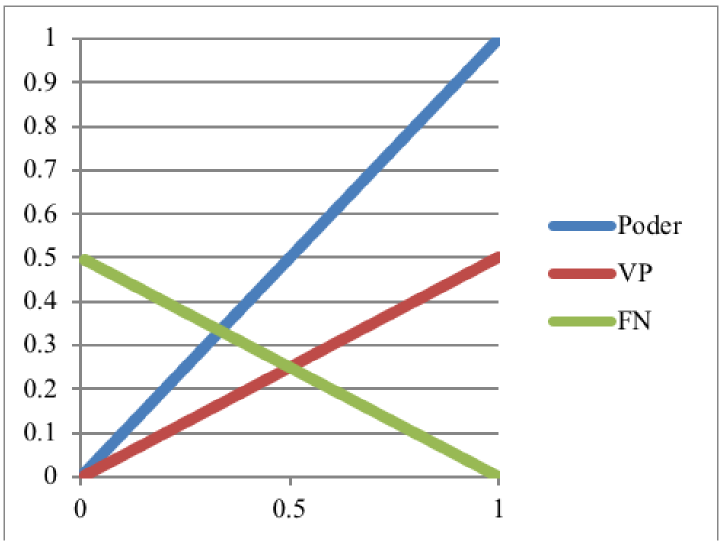
Fuente: Elaboración propia

Figura 8. La H_0 tiene 50 % de ser cierta



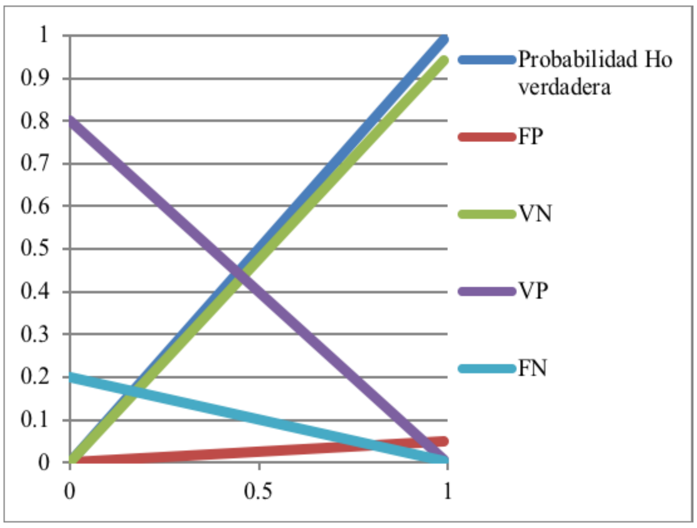
Fuente: Elaboración propia

Figura 9. La H_0 tiene 50 % de ser falsa



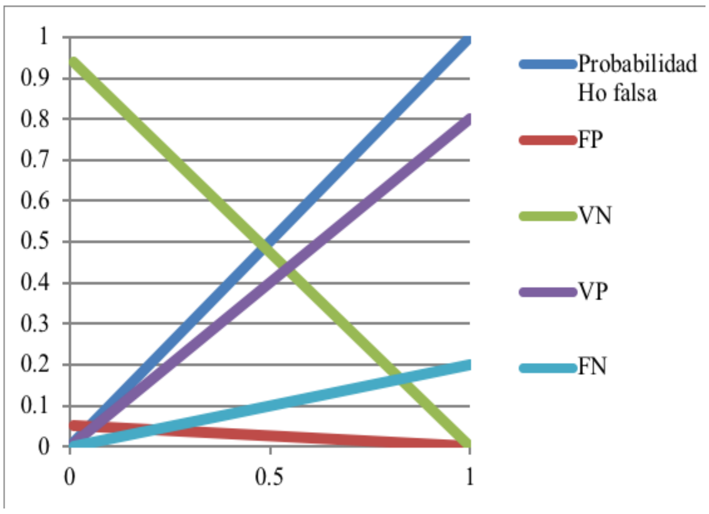
Fuente: Elaboración propia

Figura 10. Probabilidad de un H_0 verdadera



Fuente: Elaboración Propia

Figura 11. Probabilidad de una H_0 sea falsa



Fuente: Elaboración propia

Assuming that H_0 is true, this means that the population from which the control group and the treatment group came are the same. That is, they come from the same population because both distributions perfectly overlap each other. Therefore, if there is a difference between the two at the end of the experiment, it will be due to the treatment that caused this discrepancy (without considering the internal and external threats to the experiment). If the H_0 is false (there is a difference between the averages of these: true effect; eg, some treatment that improved the average), then the probability that a second study results in a true effect is equal to the power that is had in this , other things being the same (Lakens, s. f.).

Method

A quantitative, exploratory methodology was used, with descriptive and inferential statistics in order to evaluate the quality of 34 articles. For this, four criteria were used:

1) Have discussed, calculated and obtained enough power to reject a false null hypothesis.

2) Observe if the study in question is part of a series of replications.

3) See if the databases were given access for potential replicas.

4) Report statistics suggested by the APA (2001, 2020) from 2001 to the present day.

The articles in the sample were published in a series of peer-reviewed journals in the ranking of the Ibero-American Network of Innovation and Scientific Knowledge (Redib). Redib was selected because the potential audience for this study could be among researchers who have published in the journals in the sample or use them for their work. In addition, the Redib contains links to open access journals, which complies with the principle of accessibility to knowledge that is also sought in this manuscript.

In summary, the method consisted in defining certain bases for selecting a series of articles from electronic journals (see the next section). Then, these were read to capture information and statistics regarding a series of criteria to analyze their quality (the instrument consisted of a table to capture information, see table 2). Part of the analysis of the statistics of the articles was descriptive (table 3) and another part was with inferential statistics (tables 4 and 5). In the following paragraphs of the method, more details of the previous steps are given, information and statistics captured, as well as analyzes carried out.

The analysis instrument was through an Excel document (Table 2), where the various selected elements were captured, for subsequent analysis in the statistical software SPSS version 25.

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

Respuesta	*1)* Poder	*2)* Réplica	*3) Acceso*	*4) x̄*	*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 x̄ = 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

Sample

The articles in the sample (n = 34) of the present study were obtained from journals (20) that appeared in the Redib 2018 ranking: subject area (social sciences and humanities); subjects (education and educational research), and country (all countries). Only articles in Spanish and English were included because they were readable by the present authors. The keywords were: experiment (21 articles in Spanish) and experiment (13 in English). The population of this sample was 92 journals and 502 articles. The bases for selecting the articles were:

- Have carried out an experimental design in educational research: i. e., to compare groups when one of them received some treatment that differentiated it from the other, as well as a treatment received between the pre and post-test.
- Having used a parametric analysis (i. E., A t or F test) because it is recommended to infer causes and effects (Maxwell et al., 2018). Some articles also contained some non-parametric ones.

The time interval of publication of the articles was 2004-2019: specifically 23.4% covered the interval 2004-2012 and 76.5% (2013-2019). In the articles, English language learning treatments and the effect of different exam formats, among others, were carried out.

The articles in the sample were searched for information related to these four criteria, but none of the articles in the sample met any of the first three criteria (see Table 2 and the results section for more details).

For criterion four, the following data were extracted from the sample of articles to be analyzed with inferential statistics (Table 3):

- Sample size (n): all authors described the size of their study sample complying with the APA (2001, 2020).
- Averages, SD, type of analysis, test statistics (eg, calculated value t and F), degrees of freedom (df) and p. There was some variation in reporting these statistics.
- Confidence intervals (CI) and effect size. On the contrary, other statistics were the least specified: not complying with the APA (2001, 2020).

Although it was not part of the quality criteria, section B of table 3 contemplates a principle of the epistemology of science (parsimonious model): a simpler model with more power to explain and with fewer assumptions and variables is preferable to another more complicated (Sarkar and Pfeifer, 2006). In the context of the present work, several of the authors compared three averages of three independent groups with a t-test, doing the following: group two with two, one with three, and two with three (inflation). . To correct for this inflation, the α must be adjusted using the Bonferroni correction: α / number of tests (α / n) (Armstrong, 2014). Using the previous example of three t tests, the α should have been adjusted to 0.05 / 3 = 0.016. Also, a more parsimonious model can be used for the example above: the analysis of variance (Anova) of a factor that needs a comparison between the three groups. The caveat is that if you're only interested in a test, you don't need to adjust the α. Based on the articles in the sample that used several independent and dependent variables on several occasions, it is recommended to use multivariate analysis of variance (Manova) (Maxwell *et al*., 2018).

Section C of Table 3 shows the measures of central tendency of the sample size of the articles. For example, the average article size was 75.12 participants. Also, the measures of dispersion of the sample sizes are shown. An example is the standard deviation, which was 46.36 participants; A 95% confidence interval, minimum and maximum value is also shown. In addition, the measures of central tendency and variation of the number of analyzes of the articles in the sample were quantified. For example, the average number of analyzes of the articles was 14.94 for the parametric ones (SD = 26.21) and for the non-parametric ones 1.67 with a SD = 1.19.

Given the lack of a description of power in the studies, it was decided to estimate this factor with part of the information included there (although in most cases the information was not sufficient because effect sizes were lacking to calculate power). G * Power was used (free software that can help calculate n a priori to have enough power, as well as a posteriori; created and documented by Faul, Erdfelder, Buchner and Lang, 2009). In section C of table 3 the measures of central tendency and the variation of power that was estimated for the present manuscript are shown (eg, mean power = 0.74 with SD = 0.27).

A power analysis is important because it allows researchers to plan what resources will be needed to enroll or select the desired number of individuals for the study (APA, 2020). In order to describe the data, the possible normal distribution of the power coefficients was calculated. Power estimates did not pass the Kolmogorov-Smirnov test for normal distributions ($p = 0.001 < \alpha = 0.05$). Other evidence against the normality of the data was the difference between the central tendency statistics that do not coincide: mean (0.74), median (0.83) and mode (0.99): the mean and the CI are affected by an abnormal distribution. However, kurtosis (-0.679) and bias (-0.712) were within ± 2, indicating a possible normal distribution. Given this contradictory evidence, an analysis with percentages was chosen: the estimated power values above 80% were 58.8% and 41.2% below this. That is, most of the articles in the sample would have met the statistical power of 80%, other things being the same. However, it was the job of these and these authors to show that they had enough power to argue for some effect of their experimental designs.

Tabla 3. Datos descriptivos de artículos de la muestra

Sección A	Número	%	Número	%
	Encontrado	Encontrado	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 IC$_{95\ \%}$ Valor mínimo Valor máximo	46.36 [59.30, 90.94] 10 205	26.21 [13.4, 16.4] 1 120	1.19 [1.27, 2.07] 0 39	0.27 0.08 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

Inferential statistical analysis

For inferential analyzes, two types of analysis were used with the chi-square test (χ^2): *a)* goodness of fit and b) independence. This to see, with the use of SPSS version 25, if there was a difference between the expected and observed frequencies. Both tests were taken from Berenson, Levine, Szabat and Stephan (2019), Greenwood and Nikulin (1996), Hinkle, Wiersma and Jurs (2003) and Kohler (2020).

The goodness-of-fit test involves a single sample (observed frequency) that is compared to an expected frequency, which is based on a certain expectation. In this case, the expectation was that 85% of the articles would comply with reporting any of the statistics in Table 4: (i. E., Mean, SD, CI, effect size, type of analysis, test statistic, df and p). Therefore, 15% were expected not to report these aforementioned statistics. The expectation was taken from one of the assumptions of χ^2, namely: it is necessary to have at least five cases at the expected frequency to be able to use the calculated probability that is given for the test statistic χ^2 (Berenson *et al*., 2019). Here 5 of 34 articles represent 15% (the expected frequency). The hypotheses of the goodness-of-fit test are as follows:

- Null hypothesis (H0): the observed frequency = the expected frequency.
- Alternative hypothesis (HA): the observed frequency ≠ the expected frequency.

Since eight comparisons were made, the alpha (α) must be adjusted to avoid inflation of the type I error with the aforementioned Bonferroni correction. Thus, after dividing the alpha (i. E., 0.05) by the number of tests, we obtained as a result that the adjusted alpha coefficient was 0.05 / 8 = 0.0063. Therefore, the criterion for rejecting the null hypothesis was: $\alpha_{\text{ajustado}} = 0.0063$.

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

Another aspect of the goodness-of-fit test is the effect size: Cohen's W, which measures the discrepancy between pairs of proportions in cells (Cohen, 1988). The larger the difference between the expected and the observed frequency, the larger the effect size. The three different sizes are: W = 0.10 (small), W = 0.30 (medium) and W = 0.50 (large), when there is no reference in the literature.

Likewise, the χ^2 test of independence involves two variables to see how closely they are related. If they are not related: expected frequency = observed frequency: $p \geq \alpha$ and it is concluded that they are independent. On the contrary, if they are related: expected frequency ≠ observed frequency: $p < \alpha$ and it is concluded that they are dependent. On this occasion, the hypotheses were:

- Null hypothesis (H_0): the two categorical variables are independent (the observed frequency = the expected frequency).

- Alternative hypothesis (H_A): the two categorical variables are dependent (the observed frequency ≠ the expected frequency)

The criterion to reject the null hypothesis was a traditional $\alpha = 0.05$. since it is only an omnibus test with two variables (2 × 8): reported (yes or no) and statistics (eight levels: mean, SD, CI, effect size, type of analysis, test statistic, df and p). The effect size used in an independence test is Cramer's V (Akoglu, 2018): i. e., with a df = 1, a V = 0.10 (small), V = 0.30 (medium) and V = 0.50 (large).

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

On the other hand, an omnibus test indicates that $p \geq \alpha$ or $p < \alpha$, but does not indicate exactly where the difference between the expected and observed frequencies may lie. To identify it, a post hoc analysis is carried out (described by Beasley and Schumacker [1995]). Specifically, the standardized adjusted residuals are calculated in SPSS 25. Manually, these standardized adjusted residuals are squared to obtain values of $\chi^2_{\text{ calculado}}$ for each level (in this case, it's eight levels times two = eight). With these values of $\chi^2_{\text{ calculado}}$, the calculated probability (p) is calculated with a SPSS function (Beasley and Schumacker, 1995). For this post hoc test and to avoid inflation of the type I error, it is necessary to use the Bonferroni correction, the p-value at 14 tests: 0.05 / 16 = 0.0031.

Results

For descriptive results, three criteria were not met because in 100% of the publications power was not shown in the analyzes, they were not part of replications in other studies, and there was no accessibility to databases (see Table 2). One of my personal expectations was to find that most of the authors mentioned statistical power in some way in their articles. With regard to the replica element, some were expected to be repetitions of others. Likewise, it was expected that approximately 76.5% of the sample would have given access to their databases, because only this percentage corresponds to 2013 when the Center for Open Science was created: this has not been the only possibility to store databases, but it was taken as an expectation.

For the inferential results of the tests of χ^2 of goodness of fit, it was found that three of the eight tests (with a criterion to reject the null hypothesis of $\alpha_{\text{ ajustado}} = 0.0063$) they had a statistically significant difference with respect to the expectation that they did report = 29 and did not report = 5 (table 5). These were the CI (yes they reported = 6 and they did not report = 28), the effect size (yes = 12 and no = 22) and the degrees of freedom (df; yes = 22 and no = 12). Therefore, the null hypothesis is rejected in the case of these three statistics and the evidence supports the alternative that the observed frequencies are different from those expected and this is probably not random but there is an effect. The effect size W was 2.47, 1.81 and 0.74 for these three statistics respectively (table 6) and this places them in a large effect, according to Cohen (1988), which means that there is a large difference between the frequencies. It would be complex to speculate on the reason why the reporting of these statistics was omitted given that since the APA (2001) to date it has insisted on reporting them. Cumming and Calin-Jageman (2017), as well as APA (2020) are recommended for details and implications of these statistics.

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 α_{ajustado} = 0.0063.

Fuente: Elaboración propia

For the inferential results of the omnibus test of χ^2 independence, it was found that $\chi^2_{\text{calculado}}$ = 84.817, *df* = 7, *p* < 0.00001 and *V* of Cramer = 0.558 (with a criterion to reject the null hypothesis of α = .05). Therefore, the null hypothesis is rejected and it is also concluded that there is a dependency between the variables. Likewise, the effect size V was 0.558 (large) (Akoglu, 2018; Cohen, 1988), which means that there is a large relationship between the variables. However, as it is an omnibus test, it is not known where the difference lies between the expected and observed frequencies, so the post hoc test of χ^2 of Beasley and Schumacker (1995) with a α_{ajustado} = 0.0031. This last test showed that the statistically significant difference lay in the CI (yes they reported = 6 and they did not report = 28), effect size (yes = 12 and no = 22) and the type of analysis (yes = 32 and no = 2). Again, the CI and effect size reappear in the independence test, where their observed frequencies are different from those expected, they are less reported than would be expected. In contrast, the type of analysis also had a statistically significant difference, but contrary to these last two statistics, because it was the most reported statistic of the eight (see table 5). Simply put, reporting a statistic value depends on the type of statistic you are talking about. In this case, the CI and effect size were underused and the type of analysis was overused. Again, it would be difficult to speculate on the reason that led to the omission of the reporting of these statistics, given that from 2001 to date the APA has insisted on reporting them. Cumming and Calin-Jageman (2017), as well as APA (2020) are recommended for details and implications of these statistics.

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 $\alpha_{\text{ajustado}} = 0.0031$.

Fuente: Elaboración propia

Discussion

Bringing back the research question, "What has been the quality of the reporting of some statistics of refereed educational publications related to experimental processes?", And given the criteria of power, replication, accessibility and the reporting of the statistics with differences Statistically significant, the evaluation turned out to be considered of poor quality (following the aforementioned Kotz scale, 2006). For the future, publications can be greatly improved by following these criteria. The objectives were met by evaluating the statistical quality of refereed articles (n = 34) with experimental designs and listing a series of recommendations to increase their quality (APA, 2020; Cohen, 1988; Cumming and Calin-Jageman, 2017; Maxwell et al. ., 2018).

By not detailing the statistical power, the probability of having rejected a false H0 is not known. For example, with a power of 80%, by increasing the probability that an H0 is false (ie, a synonym is that the HA is true), the probability of observing a VP increases ($p < \alpha$) and the probability of a FP decreases ($p \geq \alpha$) (Harms and Lakens, 2018). With the above, and a p <alpha, it could be observed in the samples if there was a statistical effect of a treatment in an experimental design. If so, this would be promising to replicate and see if the phenomenon repeats (Feynman, 1974; Greenland et al., 2016; Harms and Lakens, 2018). To make a replica, it is necessary to have all the relevant information, so giving access to databases and analysis is a condition that must be met yes or yes, only in this way is it possible to contribute to the advancement of knowledge and knowledge. science (Carey, 2011; Cumming, G. and Calin-Jageman, 2017). Otherwise, without knowing the power or following a replication or giving access to databases and omitting relevant statistics in publications, the results of an investigation can become questionable (Cumming and Calin-Jageman, 2017). APA (2020) has recommended reporting power estimation and other methods used to determine the precision of parameter estimates. The process of determining the number of cases, participants, or observations that a study would need to reach a desired power level with a certain effect size and a certain level of significance to reject the null hypothesis.

The majority of the authors of the sample included the mean, SD, type of analysis, and p, and all declared n. In contrast, the minority calculated a df, CI, and effect size (Table 3). To see if these differences in frequencies were due to mere chance or there was an effect, two chi-square tests were carried out: goodness of fit (Table 6) and independence test (Table 7). The degrees of freedom (df) had a statistically significant difference in the goodness of fit test ($p = 0.001 <$adjusted $\alpha = 0.0063$). This statistically significant difference was evidence that the reporting of this statistic was sloppy and was probably not random. Its report is recommended because it serves to identify a critical F or t value to reject or not a null hypothesis, as well as to have enough information to replicate a study.

Two other statistics that were worrisome because they were not sufficiently reported were the CI and the effect size. The reason was that their calculated probabilities were lower than the adjusted alpha coefficients of the goodness test (for the CI and effect size their $p = 0.00001 < \alpha_{\text{ajustado}} = 0.0063$) and in the independence test (IC, $p < 0.00001$ and effect size, $p = 0.000011$, that were less than $\alpha_{\text{ajustado}} = 0.0031$). These two statistics were also neglected despite having been identified as important since 2001 by the APA. A CI helps to estimate the value of the parameter of interest (e.g., 95% means that if 100 similar samples are taken, 95 of them will contain the population parameter and five will not; and the error is also estimated, which it can be some points, if there is an error of three points, a CI of 70 points in the sample, it would be expected that the population parameter would be between 67 and 73 points). Effect sizes are often interpreted as indicative of the practical importance of a research finding, that is, the degree to which the phenomenon is present in the population or the degree to which the null hypothesis is false (APA, 2020) .

Although the authors did not touch on the issue of power, this was estimated with some of the information in the articles and assuming others for the present study. The G * Power calculator requires the following data for an a priori test (Faul et al., 2009):

- The coefficient of d (Cohen's d; see Cumming and Calin-Jageman, 2017;
- The coefficient of α;
- Desired power (at least 80%) and
- The size of the groups (n);

On the other hand, G * Power needs for the posterior test: d, α and n (Faul *et al*., 2009).

Once the power of the 34 articles in the sample was estimated, it was obtained 41.2% below and 58.8% above 80%, but these numbers have to be taken with caution because only 35.29% estimated the effect size, so assumed a median effect size for these power estimates. Similar to Ioannidis (2005), the lack of power in the estimates of these educational research articles makes questionable their conclusions to identify VP (eg, students who have truly benefited from a tutoring treatment) and FP (eg, students who falsely have benefited from a tutoring treatment). Carrying out a power analysis implies deciding a priori: a) what will be the expected effect size or what effect size will be significant based on what was found in previous research or theoretical or practical importance, b) at what level of p will the null hypothesis be rejected and c) what probability will be sufficient to reject the null hypothesis if there really is a relationship in the population (the power of the study) (APA, 2020).

Importance and significance of the results and their application and impact in the fields of knowledge

This is why the significance of the results and their application and impact in the fields of knowledge is truly important, especially in this time of COVID-19, which has touched almost every aspect of modern life, well, like this As medicine is expected to find the solution to the problem of the pandemic, the expectations of quality and performance in education must be the same and be subjected to evaluations of the highest scientific rigor. Ioannidis (2005) has long claimed in medical research that most results are false, mainly due to lack of power. Similarly, the articles in the present research sample failed to include several fundamental statistics, making the inferences made highly questionable, to say the least.

A power of 80% means that four out of five people were correctly identified as true positives when p <α. It is not enough that a p <α was found in a study, it would have to be repeated many times and published, even if it was not statistically significant. Then a metastudy could be done to see if a pattern can be observed in the phenomenon in question. (Cumming y Calin-Jageman, 2017).

Generalization

Since two statistical inference tests (goodness and independence) were carried out, the results of the present investigation could be extended to the Redib journals (20) under the following search categories: "Social Sciences and Humanities" and " Education and educational research ", in both English and Spanish. Another distinction of the articles in the sample were the keywords experiment and experiment, as well as having carried out an experimental design and a parametric analysis.

Limitations and recommendations

Among the main limitations were not having evaluated the sampling method used in the articles; not having examined the assumptions of the parametric analyzes (eg, normal distribution, missing values, linearity, homogeneity of variances and outliers, among others) (Maxwell et al., 2018), and not having covered the p-values in detail with with respect to type I error inflation, among many others. Therefore, it is recommended to cover these topics, replicate the present study and delve into the topics of effect size (Cumming y Calin-Jageman, 2017).

Journal Policy Recommendations

First of all, journals publishing articles with experimental designs are encouraged to collaborate with the Center for Open Science. This can help them publish articles with rigorous results rather than questionable results, such as those that abound in the present study sample. Another fundamental aid for manuscripts like the previous ones is the use of the APA manual (2020) and the consultation of authors such as Cohen (1988), Cumming and Calin-Jageman (2017) and Maxwell et al., (2018), among many other publications mentioned in the present study. These texts should not only be used by the authors, but also by the referees of the journals in order to support the inferences made from the results in a more solid and valid way.

Conclusions

The authors of the sample of 34 articles did not meet the criteria (stipulated by this study) of power, replication, accessibility and only partially with the reporting of some statistics. In a few words, the conclusion of the present study was that quality is considered poor by this sample of publications. These four evaluation criteria were taken from the aforementioned literature and help to make the inferences made from an experimental study more robust. According to some researchers mentioned in this study, these criteria are part of the best practices for doing empirical research both in the social sciences and in other sciences. To strengthen the inferences and advances of science itself, the following three points are recommended based on the evaluation criteria of the present study: 1) calculate the statistical power to be able to reject a false null hypothesis and to be able to minimize the number of false negatives, 2) not only seek to always carry out new studies, but to replicate existing ones to observe if any pattern is formed over time with any effect: i. e., to have evidence of some effect that is repeated and 3) not to keep the data and the analyzes, but to give accessibility to the interested parties to replicate and find potential errors and report the statistics: average, standard deviation, confidence interval, size of the effect, type of analysis, to F test statistics (among others), degrees of freedom and calculated probability. Likewise, the APA (2020) should be consulted to see what other statistics are relevant according to the type of study.

Future lines of research

It is necessary to continue analyzing the educational research literature in terms of experimental designs to see its quality. In addition, it is possible that non-experimental designs could be evaluated in terms of power level, see if they are part of a series of replications and observe the access to their data and analysis, as well as the reporting of their statistics. A possible first step would be to examine the representativeness of the study sample before a population for a possible generalization of what was found in the analyzes. That is, if the sample was taken at random from a population and has the minimum size that is marked by some formula with a confidence level (eg, 95% or 99%) and with a certain margin of error (confidence interval). When a sample is of convenience, the argument can be made that it has statistics of interest similar to the parameters of a population. This could be established with a statistical significance test to compare groups and correlate variables, as well as an effect size analysis. This would provide stronger support for generalizability of study results. Other aspects that could be reviewed from the educational research literature are the psychometric properties of the data collection instruments (exams and surveys; i. E., Validity and reliability of the scores). It is also necessary to cover the psychometric properties to see how much coherence there is between the measurements (reliability), as well as if what you want to measure (validity) is being measured. If instrument scores do not exhibit appropriate (minimum) levels of reliability and validity, there is no point in continuing with statistical analyzes and effect sizes. Finally, the present study could be replicated to evaluate the quality of the articles in other portals and journals with both experimental and non-experimental designs.

References

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