anales de psicología / annals of psychology
2024, vol. 40, nº 2 (may), 344-354
https://doi.org/10.6018/analesps.591831

ISSN online: 1695-2294.
https://revistas.um.es/analesps

Meta-analysis: Confidence intervals and Prediction intervals

Juan Botella*,1, and Julio Sánchez-Meca2

1 Universidad Autónoma de Madrid, Madrid (Spain)
2 Universidad de Murcia, Murcia (Spain)

Título: Meta-análisis: Intervalos de confianza e Intervalos de predicción.
Resumen: En los informes meta-analíticos se suelen reportar varios tipos de intervalos, hecho que ha generado cierta confusión a la hora de interpretarlos. Los intervalos de confianza reflejan la incertidumbre relacionada con un número, el tamaño del efecto medio paramétrico. Los intervalos de predicción reflejan el tamaño paramétrico probable en cualquier estudio de la misma clase que los incluidos en un meta-análisis. Su interpretación y aplicaciones son diferentes. En este artículo explicamos su diferente naturaleza y cómo se pueden utilizar para responder preguntas específicas. Se incluyen ejemplos numéricos, así como su cálculo con el paquete *metafor* en R.
Palabras clave: Intervalo de confianza. Intervalo de predicción. Meta-análisis.

Abstract: Several types of intervals are usually employed in meta-analysis, a fact that has generated some confusion when interpreting them. Confidence intervals reflect the uncertainty related to a single number, the parametric mean effect size. Prediction intervals reflect the probable parametric effect size in any study of the same class as those included in a meta-analysis. Its interpretation and applications are different. In this article we explain their different nature and how they can be used to answer specific questions. Numerical examples are included, as well as their computation with the *metafor* R package.
Keywords: Confidence interval. Prediction interval. Meta-analysis.

Introduction

In meta-analysis, several types of intervals are used, a circumstance that has generated some confusion in their interpretation, given their different nature. In this paper we explain what the two main types of intervals consist of, including numerical examples. We begin by remembering the difference between *fixed effect models* (*FEM*) and *random effects models* (*REM*), a key issue to fully understand the topic we are addressing here. We will then explain the characteristics of the two main types of intervals used in meta-analyses, but also describing an intermediate type widely used in validity generalization studies. Then we will illustrate all this with a numerical example. After highlighting the role of prediction intervals in reflecting the heterogeneity of effects we will address the discussion and a general recommendation.

For this exposition we represent the parameter that reflects the effect we are studying by θ and we assume that we have k independent estimates.

Fixed and random effects models

As we have already explained elsewhere (e.g., Botella & Sánchez-Meca, 2015), in a *FEM* (also called *common effect model*) it is assumed that the analysis being carried out refers to a single parametric value (θ). Each primary study provides an estimate of that parameter ($\hat{\theta}_i$). The estimates have a variance, which we call *sampling variance*, which must be interpreted as inaccuracy in the estimate, since it is due to the fact that each study manages data from a specific random sample, different from those of the other studies. Although it is sometimes assumed to be known (e.g., Higgins, Thompson, & Spiegelhalter, 2009), actually the sampling variance is in turn an estimated value, which we represent by $\text{var}(\hat{\theta}_i)$. As each study has a different sample size, we have a different sampling variance in each study; hence the subscript i of the sampling variances. Therefore,

a) if the sample sizes of the studies were equal, the parametric sampling variances of the studies would all be the same (although their estimates could be different), and
b) in a hypothetical case in which the sample sizes were indefinitely large, tending to infinity, the empirical variance would tend to 0.

In a *REM* it is accepted that the parametric value of each study is different. These parametric values have a distribution that is usually assumed to be normal, with mean μ_θ and variance σ^2_θ (this variance is also usually represented as τ^2 and is called *between-studies variance*, *specific variance* or *heterogeneity variance*). This means that the variance of each study's effect size estimator has two sources of variation. On the one hand, the variance of the parametric effects; on the other hand, the sampling variance of the parametric effect of each study, which is similar to the variance of the *FEM* for that particular study. Therefore,

a) if the sample sizes of the studies were equal, the parametric sampling variances of the studies would all be the same[1] (although their estimates could be different), and
b) in a hypothetical case in which the sample sizes were indefinitely large, tending to infinity, the empirical variance would tend to τ^2.

*** Correspondence address [Dirección para correspondencia]:**
Juan Botella. Universidad Autónoma de Madrid, Facultad de Psicología, Campus de Cantoblanco, c/ Ivan Pavlov, 6, 28049 Madrid (Spain).
E-mail: juan.botella@uam.es
(Article received: 6-11-2023; revised: 15-01-2024; accepted: 20-01-2024)

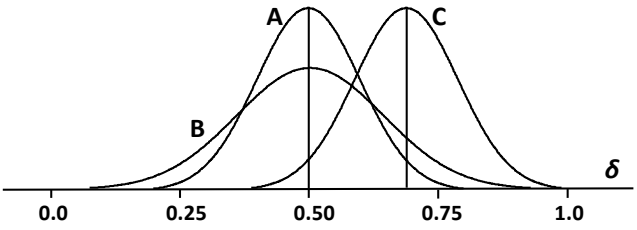
[1] As long as the value of θ_i itself is not involved in the variance, as occurs with the *standardized mean difference* (Cohen's *d*; Suero, Botella, Durán, & Blázquez-Rincón, 2023). This is not the case with other effect size indices, such as the Pearson correlation transformed with Fisher's formula.

Confidence interval (for the mean effect)

The most extended confidence interval in meta-analysis is the one that refers to the magnitude of the effect of interest. In a *FEM* the magnitude of the effect is unique: θ. Its point estimator is the weighted combination of the k estimates, $\hat{\theta}_i$, representing such combined estimate with $\hat{\theta}_\bullet$. However, in the *REM* this is not the case, rather it is assumed that there is a distribution of parametric values. A magnitude of maximum interest in this model is the estimation of the mean value of the parametric effects, μ_θ, since it is often interesting to know the "average effect" of an intervention. This magnitude is also estimated through a weighted average of the k independent estimates, $\hat{\mu}_\theta$. One of the common sources of confusion is just that in both types of models the value of interest is estimated through a weighted combination of the independent estimates provided by the k studies. But in the case of the *REM* it is about estimating what value the distribution of parametric effects is centered on. Therefore, a single, specific value is also estimated.

However, that value tells us nothing about the variations of the parametric effects. A single parametric value, μ_θ, is not sufficient to effectively describe effects that are heterogeneous (Borenstein, 2019b). Figure 1 shows three distributions of parametric effects of the *standardized mean difference* (or Cohen's *d*). Curves A and B are centered on the same value ($\mu_{\theta(A)} = \mu_{\theta(B)}$), but have different variances ($\tau^2_A < \tau^2_B$). On the contrary, curves A and C have different central values ($\mu_{\theta(C)} > \mu_{\theta(A)}$), but the same variance ($\tau^2_A = \tau^2_C$).

Figure 1
Three distributions of parametric effects



In summary, this first interval is a classic confidence interval, with which a single, specific value is estimated: the only parametric value in the *FEM* or the central value of the distribution of parametric values in the *REM*. The formulas are[2]:

$$\textit{FEM}: \ \hat{\theta}_\bullet \pm z_{\alpha/2} \cdot \hat{\sigma}_{\hat{\theta}_\bullet} \tag{1}$$

$$\textit{REM}: \ \hat{\mu}_\theta \pm z_{\alpha/2} \cdot \hat{\sigma}_{\hat{\mu}_\theta} \tag{2}$$

In the *FEM*, the parametric effect is estimated by a weighted average of the individual effects:

$$\hat{\theta}_\bullet = \frac{\sum \hat{w}_{FE,i} \cdot \hat{\theta}_i}{\sum \hat{w}_{FE,i}} \tag{3}$$

being $\hat{w}_{FE,i}$ the weighting factor of each study, which is obtained through the inverse of the sampling variance, var($\hat{\theta}_i$):

$$\hat{w}_{FE,i} = \frac{1}{\text{var}\left(\hat{\theta}_i\right)} \tag{4}$$

The standard error of the estimated effect is obtained through:

$$\hat{\sigma}_{\hat{\theta}_\bullet} = \frac{1}{\sqrt{\sum \hat{w}_{FE,i}}} \tag{5}$$

And $z_{\alpha/2}$ is the value corresponding to the $(\alpha/2)\cdot 100$th percentile (in absolute value) of the standard normal distribution.

In the *REM*, the parametric mean effect is also estimated by a weighted average of the individual effects, but the weighting factor is different:

$$\hat{\mu}_\theta = \frac{\sum \hat{w}_{RE,i} \cdot \hat{\theta}_i}{\sum \hat{w}_{RE,i}} \tag{6}$$

Being $\hat{w}_{RE,i}$ the weighting factor of each study, which is obtained by the inverse of the sum of the sampling variance, var($\hat{\theta}_i$), and the between-studies variance, $\hat{\tau}^2$:

$$\hat{w}_{RE,i} = \frac{1}{\text{var}\left(\hat{\theta}_i\right) + \hat{\tau}^2} \tag{7}$$

The between-studies variance is estimated by one of the methods proposed in the literature, for example, by restricted maximum likelihood (cf. e.g., Sánchez-Meca & Marín-Martínez, 2008; see also Suero, Botella, & Durán, 2023). The standard error of the estimated mean effect is calculated by:

$$\hat{\sigma}_{\hat{\mu}_\theta} = \frac{1}{\sqrt{\sum \hat{w}_{RE,i}}} \tag{8}$$

Its interpretation is the traditional one for this type of intervals: the CI95% provides a range of values with respect to

[2] We assume that the *standardized mean difference* is distributed according to the normal model, something we positively know is incorrect, although approximate. We do it for convenience and simplicity, but also because in the *R* package that we will use for the examples it is done like this. Actually, the distribution of Cohen's *d* is the non-central *Student*'s *t* (Suero, Botella, Durán, & Blázquez-Rincón, 2023).

which we have a 95% confidence that it includes the value of interest (θ under the *FEM*, and μ_θ under the *REM*). In other words, if we were to repeat the actions that have led us to that interval an indefinitely large number of times, under the same conditions, then approximately 95% of the intervals would include the value of interest.

The CI95% presented in equation (2) for the *REM* does not take into account the uncertainty in the estimate of the standard error of the mean effect, $\hat{\sigma}_{\hat{\mu}_\theta}$, nor of the between-studies variance, $\hat{\tau}^2$. As a consequence, the calculated interval's width tends to be underestimated. To solve this problem, Hartung and Knapp (2001; cf. also Sidik & Jonkman, 2002) proposed an alternative formula to calculate the CI95% that takes this uncertainty into account. On the one hand, the Hartung-Knapp method uses a *Student*'s *t* distribution with *k* – 1 degrees of freedom instead of the standard normal distribution. Second, it applies a correction factor to the variance of the mean effect. With *q* being the correction factor, it is obtained by:

$$q = \frac{1}{k-1}\sum \hat{w}_{RE,i}\left(\hat{\theta}_i - \hat{\mu}_\theta\right)^2 \quad (9)$$

Although unlikely, the correction factor, *q*, may be less than 1, in which case the variance by this method would be less than the original variance, resulting in narrower confidence intervals. In order to avoid this circumstance, Hartung and Knapp (2001) recommend truncating the value of *q*, so that they make $q^* = \text{máx}[1, q]$. Thus, the variance of the mean effect according to the Hartung-Knapp method is given by (Partlett & Riley, 2017):

$$\hat{\sigma}^2_{HK,\hat{\mu}_\theta} = q \cdot \hat{\sigma}^2_{\hat{\mu}_\theta} \quad (10)$$

In summary, the confidence interval for the Hartung-Knapp method is obtained by:

$$RE_{HK} : \hat{\mu}_\theta \pm t_{(k-1),\alpha/2} \cdot \sqrt{\hat{\sigma}^2_{HK,\hat{\mu}_\theta}} \quad (11)$$

It should be taken into account, however, that there are authors who do not recommend truncating the *q* value when it is less than 1. Simulation studies have shown a better adjustment to the confidence level and greater power when the Hartung-Knapp method is used without truncating than when following the authors' recommendation to truncate the value of *q* (Viechtbauer et al., 2015). Thus, the method for calculating the CI95% proposed by Hartung and Knapp does not include such truncation in the *R* package *metafor*. The meta-analysis module that incorporates version 28 of the IBM SPSS program includes both options, truncated and non-truncated (cf. IntHout, Ioannidis, & Borm, 2014; Jackson et al., 2017 for a review of alternatives in using the Hartung-Knapp method).

Prediction interval

As prediction intervals have a direct antecedent in the so-called credibility or validity intervals, we are going to explain these first and then we will explain the development that leads from validity intervals to prediction intervals. In the classic literature on the generalization of validity, the terms *credibility interval* and *validity interval* are used interchangeably (Hunter, & Schmidt, 1990). But we want to highlight that recently the use of the first term (credibility interval) has spread within the Bayesian approach (Schmid, Carlin, & Welton, 2021), although we will not dwell on that use here.

Credibility intervals were proposed by Hunter and Schmidt (1990) for a different objective than confidence intervals, within the framework of a type of meta-analysis known as *validity generalization*. It is only possible to calculate them under *REM*, but in the social and health sciences in general, and specifically in psychology, it is assumed that *REM* reflect better the scenario of the phenomena that interest us and are assumed by default. A consequence of assuming a *REM* scenario has to do with the expectation of effects in future studies. As can be seen in Figure 1, to characterize a distribution of effects, at least two magnitudes are needed: the central value or mean effect and the variance of the effects. The first refers to the mean effect of future studies of the type referred to in the population of studies. The second refers to how heterogeneous these effects are. Suppose we are talking about the impact of a therapy as reflected in a quantitative variable, compared to untreated patients on the waiting list. Each study with a randomized group design involving these two conditions would have a parametric effect belonging to a distribution of Cohen's index, with mean μ_θ and variance τ^2. Estimating μ_θ by $\hat{\mu}_\theta$ and its confidence interval, will only allow us to establish what the mean parametric effect would be in an infinitely large number of future similar studies. In other words, it allows us to conjecture where the distribution of effects is centered.

However, we are also interested in knowing what variations can be expected in these effects, since what interests us is each of the future applications. Specifically, if the parametric effects were very homogeneous around μ_θ, then decision making would be very direct and uncontroversial. But what happens if these parametric effects were very heterogeneous? It could be possible that in the next study the effect would be very large, much larger than the mean effect, but it could also be very small, or even null or negative. Depending on the type of problem and its consequences, it could be unacceptable for the intervention to have a very small effect or contrary to its efficacy. It is useful to have an idea, therefore, of what is the smallest effect that is reasonable to expect.

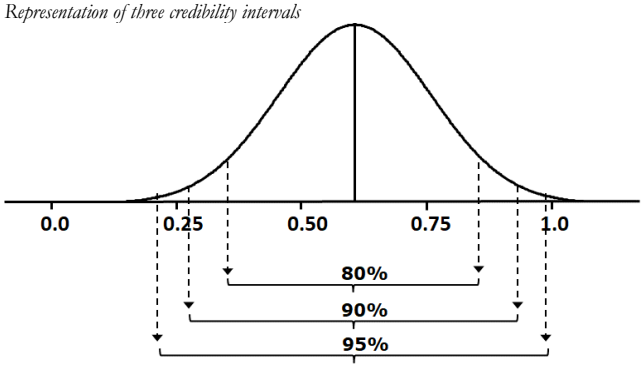
Hunter and Schmidt (1990) proposed the term *credibility intervals* for the intervals that report ranges referring to parametric values. They are obtained by:

$$\hat{\mu}_\theta \pm z_{\alpha/2} \cdot \hat{\tau} \quad (12)$$

For example, suppose that we are working with the standardized mean difference and that we estimate that the mean effect is 0.60 and the variance of these parametric effects is 0.04 (the standard deviation is 0.2). The estimated distribution assuming normality, using (12), is the one that appears in Figure 2 for three alternative confidence levels. Between the values 0.208 and 0.992 are the central 95% of the parametric values, between the values 0.271 and 0.929 are 90%, and between the values 0.344 and 0.856 are 80%. With these results, we can conclude with statements such as the following, which, as can be seen, do not refer exclusively to the average value of the distribution of effects:

a) With probability approximately .90, in a new study on an intervention of this type the effect will be between 0.271 and 0.929, and with probability .80 it will be between 0.344 and 0.856.
b) With probability approximately .95, in a new study on an intervention of this type the effect will be equal to or greater than 0.271, and with probability .90 it will be at least 0.344 (one-sided prediction intervals).

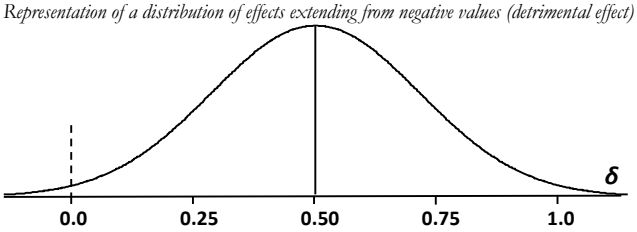
Figure 2
Representation of three credibility intervals



It is important to emphasize again that these statements do not refer to a mean value, but to the effect of a new example of this class of interventions, such as that of the next study. What is stated in example b) is important in that it provides us with a *floor value* for the effects. Sometimes it is not advisable to apply an intervention if it is not going to have a certain minimum effect, since we do not want to risk the effect being too small, or even null or negative. Statements of this type only involve the lower limit of the interval, since they refer to the effect that will be obtained, at least, with the indicated probability.

On the other hand, a highly variable distribution of effects is likely to include negative effects. For example, Figure 3 shows a distribution of effects with mean 0.50 and variance 0.09 (standard deviation 0.3). The values at two standard deviations from the central value are 0.50 ±2·0.30: [1.10; -0.10]. Then, it is possible that the effect is negative, which means that the intervention could not only not be beneficial, but could be harmful. In the example, that probability is small [P(z ≤ (0 - 0.5)/ 0.3) = .0475], but it could still be unacceptable.

Figure 3
Representation of a distribution of effects extending from negative values (detrimental effect)



In summary, credibility intervals, created in the context of the type of meta-analysis known as validity generalization, provide estimated ranges of parametric values. They reflect the probabilities that a future new study will have an effect between two values or that the effect will be at least equal to a certain value.

Let us now turn to *prediction intervals* (Higgins, Thompson, & Spiegelhalter, 2009; Riley, Higgins, & Deeks, 2011). They are conceptually identical to those of credibility and can be considered a development and sophistication of these. However, while credibility intervals are hardly used outside the scope of validity generalization studies, prediction intervals are used in very varied fields and their presence is increasing in meta-analyses of both psychology and other fields, such as medicine in general.

The objective of prediction intervals is the same as that of credibility intervals: to offer a range of probable values for the parametric effect of a future new study of the same type of those that have been included in the meta-analysis. The differences with the credibility intervals are rather technical: they assume and include in the model the uncertainty associated with the estimation of both μ_θ and τ^2. In the credibility interval, when calculating $\hat{\mu}_\theta \pm z_{\alpha/2} \cdot \hat{\tau}$ it is assumed that both values are known. Recognizing that they are estimates and, therefore, imprecise values, the uncertainty derived from the estimation process is taken into account. Specifically, if we assume that the central value of the distribution is in the range provided by its confidence interval, then the smallest value of the effect we want to identify will be at certain distance from its lower limit, not from the central value of the interval. The same goes for the upper limit of the interval. On the other hand, by recognizing the uncertainty associated with these two magnitudes, the distribution model is no longer normal, but *Student*'s *t* with (*k*-2) degrees of freedom. There is some controversy regarding the appropriate degrees of freedom, but many authors follow the suggestion of Higgins et al (2009) to use the t_{k-2} distribution as a reasonable and practical option (e.g., Borenstein et al, 2021; Stijnen, White, & Schmid, 2021).

In summary, the prediction interval can be obtained through (Higgins et al, 2009),

$$\hat{\mu}_\theta \pm {}_{\alpha/2}t_{k-2} \cdot \sqrt{\hat{\tau}^2 + \hat{\sigma}^2_{\hat{\mu}_\theta}} \quad (13)$$

The prediction interval defined in (13) uses the sample variance originally proposed for the mean effect in a *REM*, $\hat{\sigma}^2_{\hat{\mu}_\theta}$, which is calculated by the square of equation (8). Instead, it is more advisable to use the variance proposed by Hartung and Knapp, $\hat{\sigma}^2_{HK,\hat{\mu}_\theta}$, defined in equation (10), which takes into account the uncertainty when estimating both the between-studies variance and the mean effect (Partlett & Riley, 2017):

$$\hat{\mu}_\theta \pm {}_{\alpha/2}t_{k-2} \cdot \sqrt{\hat{\tau}^2 + \hat{\sigma}^2_{HK,\hat{\mu}_\theta}} \quad (14)$$

An important difference between confidence intervals and prediction intervals is that, if the number of studies is progressively increased, the confidence interval reduces its width and tends to 0 when the number of studies tends to infinity. This is because the only component of the variance involved, $\hat{\sigma}^2_{\hat{\mu}_\theta}$, decreases as the number of studies increases. On the contrary, no matter how many studies are added, the width of the prediction interval has a floor value that cannot be exceeded. This is because one of the components of the variance involved, τ^2, does not change with the number of studies (although its estimate does improve).

An example

In this section we illustrate what has been said so far through a numerical example (the *R* code in the appendix allows reproducing the calculations). The following table shows the *standardized mean difference* estimates (*g*, or corrected Cohen's *d*) from 15 independent studies and their sample sizes. With these three values we have obtained the sampling variances that appear in the last column with the approximate formula of Hedges and Olkin (1985; formula [2.8] in Botella & Sánchez-Meca, 2015). The results offered by *metafor* (Viechtbauer, 2010) also appear in the table when fitting the *FEM* and the *REM*, estimating the specific variance by restricted maximum likelihood. We have numerous alternative methods to estimate τ^2, but there is still not enough agreement on the best options (Blázquez-Rincón, Sánchez-Meca, Botella & Suero, 2023; Langan, et al, 2017, 2019; Veroniki et al, 2016). For this example, we have chosen a frequently used and recommended method, restricted maximum likelihood, that does not have the disadvantages of the well-known method of moments (Viechtbauer, 2005). We detail in the right part of the Table 1 the calculations for the types of intervals explained above.

Table 1
Calculations for the types of intervals

Study	N1	N2	*g*	Var_g
1	11	11	0.2700	0.18350
2	123	135	0.2991	0.01571
3	14	16	0.0584	0.13399
4	36	35	-0.0198	0.05635
5	20	20	0.7645	0.10731
6	24	23	0.9341	0.09443
7	20	12	0.1560	0.13371
8	80	75	0.2388	0.02602
9	18	18	0.0293	0.11112
10	16	16	1.2087	0.14783
11	32	20	0.4728	0.08340
12	16	18	0.7811	0.12703
13	20	20	1.0977	0.11506
14	76	58	0.8551	0.03313
15	24	16	0.2156	0.10475

Fixed-Effects Model (k = 15)
I^2 (total heterogeneity / total variability): 49.74%
H^2 (total variability / sampling variability): 1.99
Test for Heterogeneity:
Q(df = 14) = 27.8570, *p*-val = .0149
Model Results:

estimate	se	zval	pval	ci.lb	ci.ub
0.4326	0.0640	6.7593	< .0001	0.3071	0.5580

Random-Effects Model (k = 15; tau^2 estimator: REML)
tau^2 (estimated amount of total heterogeneity): .0690 (SE = .0551)
tau (square root of estimated tau^2 value): .2627
I^2 (total heterogeneity / total variability): 51.24%
H^2 (total variability / sampling variability): 2.05
Test for Heterogeneity:
Q(df = 14) = 27.8570, *p*-val = .0149
Model Results:

estimate	se	zval	pval	ci.lb	ci.ub
0.4707	0.1007	4.6760	< .0001	0.2734	0.6680***

Model Results (Hartung-Knapp method):

estimate	se	tval	df	pval	ci.lb	ci.ub
0.4707	0.1012	4.6493	14	.0004	0.2535	0.6878***

95% Prediction Interval (assuming a standard normal distr.):

pred	se	ci.lb	ci.ub	pi.lb	pi.ub
0.4707	0.1007	0.2734	0.6680	-0.0808	1.0221

95% Prediction Interval (Hartung-Knapp method):

pred	se	ci.lb	ci.ub	pi.lb	pi.ub
0.4707	0.1012	0.2535	0.6878	-0.1332	1.0746

The *confidence interval* is the only one that can be calculated under a *FEM*. The estimates for both models are those provided directly by *metafor*, which we detail below applying formulas (1), (2) and (11):

CI95%(*FEM*): 0.4326 ± 1.96·0.0640: **[0.307; 0.558]**
CI95%(*REM*): 0.4707 ± 1.96·0.1007: **[0.273; 0.668]**
CI95%(*REM* by HK): 0.4707 ± 2.145·0.1012: **[0.254; 0.688]**

The value 2.145 corresponds to the 97.5th percentile of the *Student t* distribution with k - 1 = 15 - 1 = 14 degrees of freedom. The value 0.1012 is the square root of the magnitude obtained with equation (10).

As we have already explained, the credibility and prediction intervals can only be calculated under the *REM*. Specifically, the credibility or validity interval would be the one provided by formula (12):

VI95%: 0.4707 ± 1.96·0.2627: **[-0.044; 0.986]**

The prediction intervals are those provided by formulas (13) and (14) (the value 2.16 corresponds to the 97.5th percentile in the *t* distribution with k-2 = 13 degrees of freedom):

PI95%: 0.4707 ± 2.16·sqrt(0.2627² + 0.1007²) =
= 0.4707 ± 2.16·0.2813: **[-0.137; 1.078]**
PI95% by HK: .4707 ± 2.16·sqrt(0.2627² + 0.01024) =
= 0.4707 ± 2.16·0.2813: **[-0.137; 1.079]**

Note that the PI95% calculated with equation (13), [-0.137; 1.078], does not match the one reported in the output of the *metafor* program: PI95%[-0.0808; 1.0221]. This is because the formula implemented in *metafor* is not equation (13), but rather assumes a standard normal distribution instead of a *Student*'s *t* distribution. The formula implemented in *metafor* is, therefore (Viechtbauer, 2023, p. 32):

$$\hat{\mu}_\theta \pm z_{\alpha/2} \cdot \sqrt{\hat{\tau}^2 + \hat{\sigma}^2_{\hat{\mu}_\theta}} \qquad (15)$$

Likewise, note that the PI95% by HK calculated with equation (14), [-0.137; 1.079], also does not match the one reported in the output of *metafor*: PI95%HK[-0.1332; 1.0746]. This is because *metafor* uses (k – 1) instead of (k – 2) as degrees of freedom of the *Student*'s *t* distribution. Thus, the formula implemented in *metafor* for the PI95% by the Hartung-Knapp method is (Viechtbauer, 2023, p. 32):

$$\hat{\mu}_\theta \pm {}_{\alpha/2}t_{k-1} \cdot \sqrt{\hat{\tau}^2 + \hat{\sigma}^2_{HK,\hat{\mu}_\theta}} \qquad (16)$$

Note that the prediction interval is more imprecise (wider) than that of validity or credibility and, in turn, the Hartung-Knapp prediction interval is wider than the one that does not apply this method. The results can be interpreted in the following way. The CI95% indicates, under the *FEM*, that we can conclude, with 95% confidence, that the parametric (single) value involved is in the range 0.307 – 0.558. Under the *REM*, the 95%CI indicates that we can conclude, with 95% confidence, that the mean value of the effects in future studies of this type will tend to be a value in the range 0.273 – 0.668.

Regarding the PI95%, it is interpreted concluding that, with 95% confidence, the effect of a future study of this type will be in the range -0.137 – 1.078, or in the range -0.137 – 1.079, depending on whether or not we have used the Hartung-Knapp method. Note the difference between this conclusion and that of the CI95%. In the PI95% it is concluded *about the effect of a future intervention*, while in the CI95% it is concluded *about the average effect of future interventions*.

When it is important not to implement or recommend interventions with effects that might be too small (or, of course, interventions that might be harmful), then it is appropriate to provide interpretations in that sense. We have already seen in the example in Figure 2 that we can do it using one-sided intervals. For example, in this case we can conclude that the effect of the next intervention of this type will be, with 97.5% confidence, greater than -0.137. We can also calculate and interpret other values, for example in the following ways:

a) The effect of a future intervention of this type will be, with 95% confidence, greater than (-1.771 is the value corresponding to the 5th percentile of the *t* distribution with 13 degrees of freedom):

$$0.4707 - 1.771 \cdot 0.2813 = -0.027$$

b) It can be estimated the probability of a positive effect of the next intervention. To do this, we calculate the value of the statistic as:

$$(0 - 0.4707) \,/\, 0.2813 = -1.6733$$

As in the *Student t* distribution with 13 degrees of freedom this value has a cumulative probability of .059, the desired probability is 1 – .059 = .941.

c) If we consider that there is a minimum effect size, essential for it to be reasonable to apply the intervention, we can estimate the probability that such value will be reached and thus make better decisions. Suppose that in this example the minimum size is $\delta = 0.25$. We estimate the probability that the effect is at least that magnitude as follows. First we calculate the value of the statistic for that effect value:

$$(0.25 - 0.4707) \,/\, 0.2813 = -0.7846$$

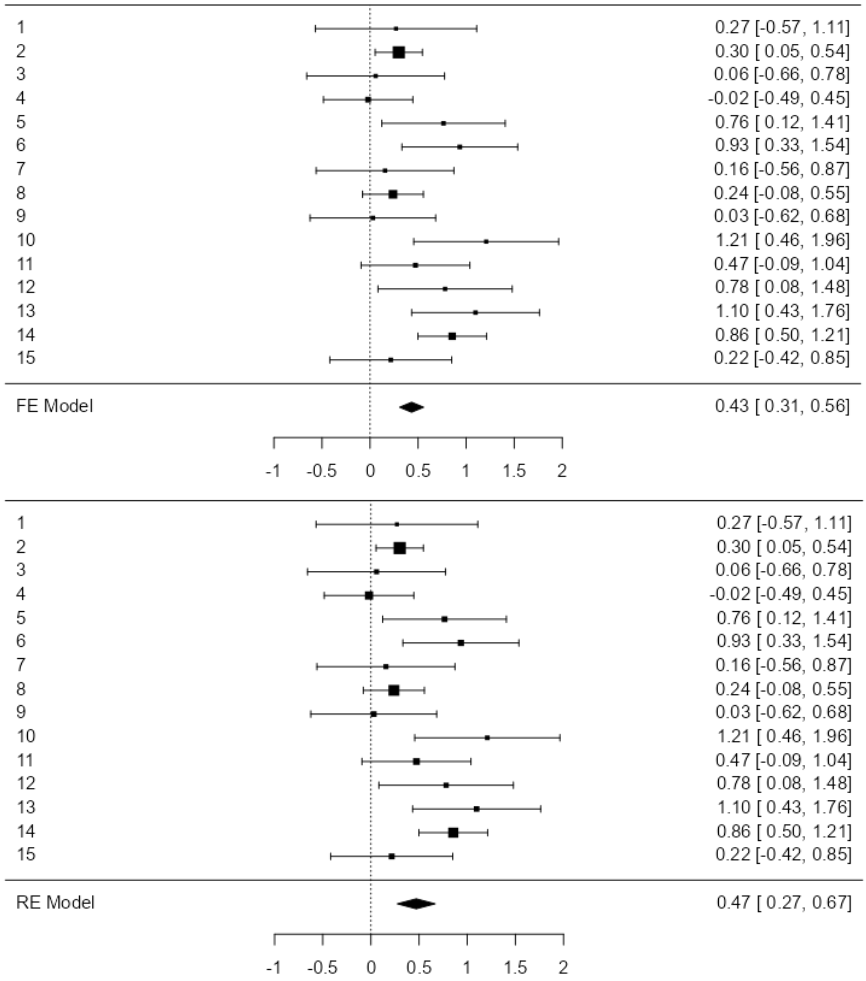
In the *Student*'s *t* distribution with 13 degrees of freedom this value has a cumulative probability of .2234. Therefore, the desired probability is 1 – .2234 = .7766. In summary, the estimated probability that the effect is at least $\delta = 0.25$ is equal to .7766.

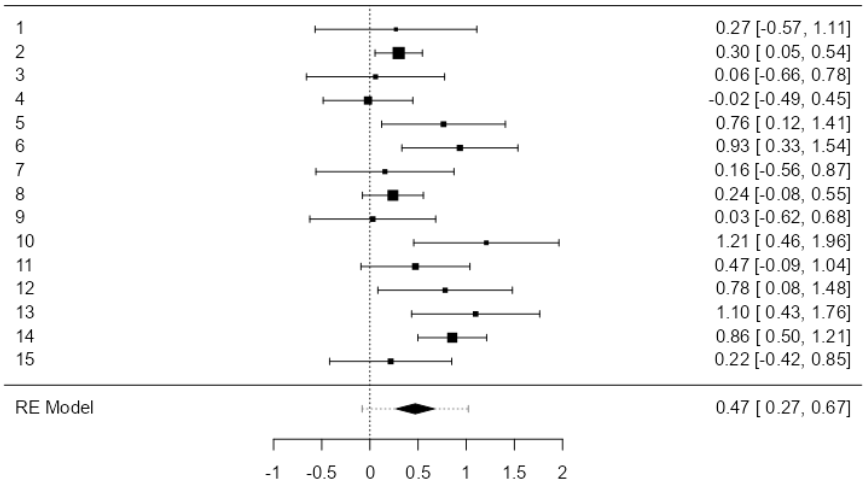
Analyzes such as those just described may be especially appropriate when studying the role of moderators, especially categorical moderators. It is possible that under one of the categories the established minimum is very likely to be reached, while in another the probability is very low. Prediction intervals can help make better decisions in the context of evidence-based practice.

Three *forest plot* graphs are presented in Figure 4. All three show in their central part the estimates provided by the individual studies in the previous example, with identical confidence intervals at the study level. The difference between them is in the combined estimates they provide and their ranges. The first one presents the combined estimate under a *FEM* (the rhombus or diamond that appears below the last study; its vertical axis represents the combined value and the ends of its horizontal axis the limits of the CI95%). In the second and third, the rhombus also represents the CI95%, but under a *REM* (without the Hartung-Knapp adjustment). Naturally, the CI95% under a *REM* is wider than under a *FEM*. In the third one, a dashed horizontal segment has been added to the rhombus, which represents the PI95%. In this third *forest plot* the different interpretation of the components is more easily noted. While the rhombus shows the range in which the central value of the effects would be in 95% of the cases, the segment shows the central range in which the parametric effect will be in 95% of the future individual studies belonging to the population of studies referred to in this meta-analysis. The amplitude of the rhombus is often incorrectly interpreted as reflecting the range of parametric effects. This error has sometimes been pointed out as one of the common or typical errors in meta-analysis (Borenstein, 2019a). Note that while the confidence interval is obtained with the *standard error* of the central value estimator, the prediction interval is obtained with the *standard deviation* of the parametric effects.

Figure 4

***Forest plots** with the studies in the example. The first with the results of a FEM. The second and third with a **REM**. The third includes the representation of the prediction interval as calculated by **metafor** with equation (15).*





The prediction interval as an index of heterogeneity

We already know the importance of reporting some estimate of the dispersion of effects in meta-analyses. Reporting the estimate of the mean effect size is not enough, even if the estimate is very precise (narrow confidence interval). It is essential to accompany this estimate with some indicator of how heterogeneous the effects are. Especially in applied fields, good decisions can only be made if the expected variations in the effect are anticipated (Borenstein, 2019b). Several ways have been proposed to reflect this dispersion, but not all of them are correct nor are those that are correct equally effective. IntHout et al (2016) focus on this issue, reviewing the most used indices and evaluating their advantages and disadvantages. Finally, they propose using the prediction interval for this objective. We refer the interested reader to that source, while here we summarize the main conclusions.

The specific variance, τ^2, is not an appropriate index because it is not in the natural metric of the effect studied, while its square root may be, but only if the values have not been transformed for the calculations (e.g., logarithmic transformation of *OR* or Fisher's *Z* for correlations). The I^2 index also does not reflect well the intended construct, since it is an index of the relative proportion of variation due to parametric effects (Higgins, & Thompson, 2002). As a consequence, a set of studies with very high sample sizes can give rise to a very high I^2 index, even if there is not large heterogeneity among the effect sizes; and vice versa, a set of studies with small sample sizes with the same between-studies variance as the previous case, can give place to a low I^2 index (Borenstein et al, 2017). Of course, the Q statistic is not useful in this sense either, since it is only a statistic to test whether the null hypothesis H_0: $\tau^2 = 0$ can be hold (Huedo-Medina et al, 2006). We have already seen that the confidence interval should not be used to make this type of interpretation, since it only reflects the uncertainty in the estimate of the average effect, not the variations in the effects. On the contrary, IntHout et al (2016) propose the use of the *prediction interval* to evaluate and communicate the degree of dispersion of the true effects. The prediction interval reflects exactly what is intended and is in the metric of the effect size index used. IntHout et al (2016) recommend, as we also do, routinely reporting prediction intervals in meta-analyses. It should be taken into account, however, that if the meta-analytic synthesis has been carried out after transforming the effect sizes of the studies (e.g., Fisher's *Z* for correlation coefficients, the natural logarithm of the odds ratio or of the risk ratio, or the logit function applied on prevalence), then the lower and upper limits of the prediction interval must be back-transformed to return them to the metric of the effect size index of interest (e.g., transforming them to coefficients correlation, odds ratios, risk ratios or prevalence). One should also be aware that the prediction intervals are obtained by estimating τ^2 and that these estimates are very imprecise with a low number of studies. Consequently, the interpretation of prediction intervals should also be taken with caution when they have been obtained with few studies (or even avoided with a very low number).

Discussion and conclusions

One of the main objectives of meta-analysis is the analysis of results oriented towards the synthesis of evidence. This synthesis serves to better understand the phenomena and the inter-relationships between the variables under study. But it should also serve to make better decisions in new research and new interventions. Prediction based on ac-

quired knowledge (or evidence-based practice) reflects better than anything else the role of meta-analysis in the applied world.

In the *REM* a certain distribution model of the parametric effects, the normal model, has been assumed. This assumption may not be appropriate, but it is assumed for convenience, because other models are also arbitrary and because when the number of studies is reasonably high the approximation is quite good. Higgins et al (2009) discuss other distributions within the frequentist approach, as well as some alternatives within the Bayesian approach.

In this article we have described what the prediction intervals are, distinguishing them both from confidence intervals (which refer to a different concept) and from their predecessor, credibility or validity intervals. We have highlighted the differences in their calculation and, above all, in their interpretation. We believe, following IntHout et al (2016) that prediction intervals should be reported much more frequently, and routinely when it comes to studies on interventions, a recommendation that is also made in other sources (e.g., Borenstein, 2019a, 2019b; Borenstein et al, 2017, 2021; Higgins et al, 2019; Schmid et al, 2021).

Complementary information

Financial support.- This research was supported by the Ministerio de Ciencia e Innovación of Spain (project reference: PID2021-122404NB-I00). Contact: Juan Botella.
Conflict of interest.- The authors declare no conflict of interest.

References

Blázquez-Rincón, D., Sánchez-Meca, J., Botella, J., & Suero, M. (2023). Heterogeneity estimation in meta-analysis of standardized mean differences when the distribution of random effects departs from normal: A Monte Carlo simulation study. *BMC Medical Research Methodology*, 23(1), 19. https://doi.org/1 .1186/s12874-022-01809-0

Borenstein, M. (2019a). *Common mistakes in meta-analysis*. Englewood, NJ: Biostat inc.

Borenstein, M. (2019b). Heterogeneity in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (eds.), *The Handbook of Research Synthesis and Meta-analysis*, 3rd ed. (pp. 453-468). New York: Russell Sage Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5-18. https://doi.org/1 .1002/jrsm.1230

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2ª ed.). Chichester, UK: John Wiley and sons. [Chapter 17]

Botella, J., & Sánchez-Meca, J. (2015). [Meta-analysis in Social and Health Sciences] *Meta-análisis en Ciencias Sociales y de la Salud*. Madrid: Editorial Síntesis.

Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine, 20*(12), 1771-1782. https://doi.org/1 .1002/sim.791

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-analysis*. San Diego, CA: Academic Press.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds) (2019). *Cochrane handbook for systematic reviews of interventions*. (2nd edition). Wiley.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11), 1539-1558. https://doi.org/1 .1002/sim.1186

Higgins, J. P., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1), 137-159. https://doi.org/1 .1111/j.1467-985X.2008.00552.x

Huedo-Medina, T., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistics or I^2 index? *Psychological Methods*, 11, 193–206. https://doi.org/1 .1037/1082-989X.11.2.193

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

IntHout, J., Ioannidis, J. P., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology, 14*(25). http://www.biomedcentral.com/1471-2288/14/25

IntHout, J., Ioannidis, J. P., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ open*, *6*(7), e010247. http://doi:1 .1136/bmjopen-2015-010247

Jackson, D., Law, M., Rücker, G., & Schwarzer, G. (2017). The Hartung-Knapp modification for random-effects meta-analysis: A useful refinement but are there any residual concerns? *Statistics in Medicine, 36*, 3923–3934. https://doi.org/1 .1002/sim.7411

Langan, D., Higgins, J. P., & Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Research Synthesis Methods*, 8(2), 181-198. https://doi.org/1 .1002/jrsm.1198

Langan, D., Higgins, J. P., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., ... & Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*, 10(1), 83-98. https://doi.org/1 .1002/jrsm.1316

Partlett, C., & Riley, R.D. (2017). Random effects meta-analysis: Coverage performance of 95% confidence and prediction intervals following REML estimation. *Statistics in Medicine, 36*, 301-317. https://doi.org/1 .1002/sim.7140

Riley, R. D., Higgins, J. P., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ*, *342*. https://doi.org/1 .1136/bmj.d549

Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods, 13*, 31-48. https://doi.org/1 .1037/1082-989X.13.1.31

Schmid, C. H., Stijnen, T., & White, I. (Eds.). (2021). *Handbook of Meta-analysis*. CRC Press.

Schmid, C. H, Carlin, B. P., & Welton, N. J. (2021). Bayesian Methods for Meta-analysis. En *Handbook of Meta-Analysis* (pp. 41-64). Chapman and Hall/CRC.

Sidik, K., & Jonkman, J. N. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine, 21*(21), 3153–3159. https://doi.org/1 .1002/sim.1262

Stijnen, T., White, I. R., & Schmid, C. H. (2021). Analysis of univariate study-level summary data using normal models. In *Handbook of Meta-Analysis* (pp. 41-64). Chapman and Hall/CRC. [Section 4.4.4.2]

Suero, M., Botella, J., & Durán, J. I. (2023). Methods for estimating the sampling variance of the standardized mean difference. *Psychological Methods*, 28(4), 895-904. https://doi.org/1 .1037/met0000446

Suero, M., Botella, J., Durán, J. I., & Blázquez-Rincón, D. (2023, September 12). Reformulating the meta-analytical random effects model as a mixture model. https://doi.org/1 .17605/OSF.IO/V2FDE

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., ... & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1), 55-79. https://doi.org/1 .1002/jrsm.1164

Viechtbauer W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects Model. *Journal of Educational and Behavioral Statistics*, 30, 261–293. https://doi.org/1 .3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48. https://doi.org/1 .18637/jss.v036.i03

Viechtbauer, W. (2023). *Package 'metafor'*. Unpublished document. University de Maastricht.

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods, 20*, 360-374. https://doi.org/1 .1037/met0000023

Appendix

R code used for the example in the text. Before running it, an SPSS file (.sav) must be created with the columns that appear in the left panel of the example table. The columns of the effect sizes and their variances are in the columns called "g" and "Var_g", respectively.

```
#EXAMPLE OF THE TEXT, WITH 15 STUDIES
#Reading the data, in the file SPSS "Example_15_studies.sav"
library(foreign)    #The package "foreign" must be already installed
Data <- read.spss("Example_15_studies.sav ")

#Loading the package "metafor"
library("metafor")

#Fixed effect model
resFE <- rma.uni(yi=g, vi=Var_g, data=Data, method="FE")
resFE

#Random effects model, non-adjusted
resRE <- rma.uni(yi=g, vi=Var_g, data=Data, method="REML")
resRE

#Random effects model, with Knapp-Hartung adjustment
resRE_KH <- rma.uni(yi=g, vi=Var_g, data=Data, method="REML", test="knha")
resRE_KH

#Prediction interval, Random effects model non-adjusted
predict(resRE)

#Prediction interval, Random effects model with Knapp-Hartung adjustment
predict(resRE_KH)

#The three forest plot in figure 4
forest(resFE)
forest(resRE)
forest(resRE, addpred=TRUE)
```