Overview

The analysis is based on seventeen studies. The effect size index is the standardized difference in means (d).

[Indicate what an effect on either side of zero represents - for example, if it favors the treated group or the control group.]

Statistical model

The random-effects model was employed for the analysis. The studies in the analysis are assumed to be a random sample from a universe of potential studies, and this analysis will be used to make an inference to that universe.

(Borenstein, 2019; Borenstein et al., 2010; Borenstein et al., 2021; Hedges & Vevea, 1998; Higgins & Thomas, 2019)

Notes on the statistical model

This model applies when there is a universe of populations that we care about and we can assume that the studies in the analysis are a random sample of those populations. The results of our analysis will be generalized to that universe.

In general, when the studies in the analysis are pulled from the literature, this is the model we should be using. This is the model that allows us to generalize to the universe of comparable studies, which is typically what we intend to do.

At the same time, we need to be aware of the limitations of this model when employed in this way. In particular, for the model to work properly we need a sufficient number of studies to yield a reliable estimate of tau-squared, the between-study variance. Additionally, while we can say that the results apply to the universe of comparable studies, it may not be clear what that universe includes.

What is the mean effect size?

The mean effect size is 0.506 with a 95% confidence interval of 0.362 to 0.651. The mean effect size in the universe of comparable studies could fall anywhere in this interval.

The Z-value tests the null hypothesis that the mean effect size is zero. The Z-value is 6.857 with $p < 0.001$. Using a criterion alpha of 0.050, we reject the null hypothesis and conclude that in the universe of populations comparable to those in the analysis, the mean effect size is not precisely zero.

Notes on the summary effect size

The summary effect size is referred to as the mean effect size (rather than the common effect size) because we are using the random-effects model. Under this model, the true effect size varies from study to study and we are estimating the mean of these effects.

The Q-test for heterogeneity

The Q-statistic provides a test of the null hypothesis that all studies in the analysis share a common effect size. If all studies shared the same true effect size, the expected value of Q would be equal to the degrees of freedom (the number of studies minus 1). The Q-value is 30.106 with 16 degrees of freedom and $p = 0.017$. Using a criterion alpha of 0.100, we can reject the null hypothesis that the true effect size is the same in all these studies.

Notes on the Q-test

The Q-test is intended to address the question "Is there evidence that the true effect size varies (at all) across studies". The criterion alpha for the Q-test is typically set at 0.100. The rationale for this is that the test often has low statistical power, and that using a criterion of 0.100 makes it more likely that we will be able to reject the null hypothesis and prove that the effect size varies across studies.

In this case the Q-value is 30.106 with 16 degrees of freedom and $p = 0.017$. Using a criterion alpha of 0.100 we can reject the null hypothesis that the true effect size is the same in all these studies.

While we do have empirical evidence in this case that the effect size varies across studies, this is not needed to employ the random-effects model. If logic tells us that the effect size varies across studies, we should assume that it does vary even if the test for heterogeneity is not statistically significant.

Researchers sometimes assume that the Q-statistic or the corresponding p-value tell us something about the amount of variation in effects. This is a mistake. The Q-value and p-value only address the question "Is there ANY variation in effects?" Typically, the question we really care about is "HOW MUCH does the effect size vary?" This question is addressed by the prediction interval, as discussed below.

The I-squared statistic

The I-squared statistic is 47%, which tells us that some 47% of the variance in observed effects reflects variance in true effects rather than sampling error.

Notes on I-squared

There is a common belief that I-squared tells us how much the effect size varies across studies. In some cases, I-squared values have been employed to classify the amount of variation as being low, moderate, or high.

While this interpretation of I-squared is ubiquitous in some fields, it is nevertheless incorrect. I-squared is a proportion, not an absolute amount. It tells us what proportion of the variance in observed effects reflects variance in true effects rather than sampling error. It does not tell us how much the effects actually vary.

The statistic that tells us how much the effect size varies is the prediction interval, which is discussed below.

Tau-squared and tau

Tau-squared, the variance of true effect sizes, is 0.039 in d units. Tau, the standard deviation of true effect sizes, is 0.197 in d units.

The prediction interval

If we assume that the true effects are normally distributed (in d units), we can estimate that the prediction interval is 0.058 to 0.954. The true effect size in 95% of all comparable populations falls in this interval.

(Borenstein, 2019, 2020; Borenstein et al., 2021; Borenstein et al., 2017; DerSimonian & Laird, 1986, 2015; Higgins, 2008; Higgins & Thompson, 2002; Higgins et al., 2003; Higgins & Thomas, 2019; IntHout et al., 2016.)

Computations were carried out using Comprehensive Meta-Analysis Version 4 (Borenstein et. al., 2022)

Notes on the prediction interval

The prediction interval will be accurate if all relevant assumptions are met. However, if our estimate of the mean is incorrect, if our estimate of the standard deviation is incorrect, and/or the effects do not follow a normal distribution, the interval may be misleading.

Researchers are sometimes confused by the array of statistics that are reported for heterogeneity. In fact, all of these statistics are based on the same numbers - they simply focus on different aspects of heterogeneity. We need to be clear about what question we intend to ask, and then use the statistic that addresses that question.

If we want to ask, "Is there evidence of ANY variation in effect size" we should look to the Q-test.

If we want to ask, "Of the variance that we see in the observed effects, what proportion reflects variation in true effects rather than sampling error?" we should look to I-squared.

If we want to ask, "What is the variance of true effects?", we should look to Tau-squared.

If we want to ask, "In what interval do the true effects fall?", we should look to the prediction interval. This gives us the dispersion on the same scale as the mean effect size.

These are all valid statistics, but they are not interchangeable. We need to be clear about what question we intend to ask, and then use the appropriate statistic.

In the vast majority of cases, we intend to ask, "How much do the true effects vary?", and the statistic that addresses that question is the prediction interval.

One function of the prediction interval is to provide context for understanding the mean effect size.

If the prediction interval tells us that the effect size is consistent across studies, the impact of the intervention will be similar for all relevant populations, and it would probably be useful to focus on the mean effect size.

By contrast, if the prediction interval tells us that the effect size varies substantially across populations, the mean becomes less useful as a summary statistic. Rather, we would want to understand the range of effects. The prediction interval may allow us to report, that (a) the effect size is always helpful, but varies from a trivial to a moderate effect, or (b) the effect size is always helpful, and varies from a moderate to a large effect, or (c) the effect is helpful in some cases but may be harmful in others.

The accuracy of the prediction interval depends on several assumptions, including the assumption that the effects are normally distributed about the mean. Therefore, if the prediction interval includes effects that indicate an intervention is harmful it would be important to see if there are actually observed effects that support this conclusion.

Disclaimer

This synopsis is intended to assist the reviewer in interpreting the statistics. No warranty in this synopsis is expressed or implied.

In any event, a correct interpretation of these statistics is only the first step in understanding the results. The next step is to use these statistics to synthesize the results and put them in context, based on an understanding of the methods as well as the subject matter.

Learn more

This was intended only as an overview of these issues. Click Help for access to PDFs and videos that explore these issues in more detail.

References

Borenstein, M. (2019). Common Mistakes in Meta-Analysis and How to Avoid Them. Biostat, Inc.

Borenstein, M. (2020). Research Note: In a meta-analysis, the I2 index does not tell us how much the effect size varies across studies. J Physiother, 66(2), 135-139. https://doi.org/10.1016/j.jphys.2020.02.011

Borenstein, M., Hedges, L. E., Higgins, J. P. T., & Rothstein, H. R. (2022). Comprehensive Meta-Analysis Version 4. In Biostat, Inc. www.Meta-Analysis.com

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods, 1(2), 97-111. https://doi.org/10.1002/jrsm.12

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). Introduction to Meta-Analysis (Second ed.). Wiley.

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. Res Synth Methods, 8(1), 5-18. https://doi.org/10.1002/jrsm.1230

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. Control Clin Trials, 7(3), 177-188. http://www.ncbi.nlm.nih.gov/pubmed/3802833

DerSimonian, R., & Laird, N. (2015). Meta-analysis in clinical trials revisited. Contemp Clin Trials, 45(Pt A), 139-145. https://doi.org/10.1016/j.cct.2015.09.002

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Academic Press. Publisher description http://www.loc.gov/catdir/description/els032/84012469.html

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random-effects models in meta-analysis. Psychological Methods, 3(4), 486-504.

Higgins, J. P. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. Int J Epidemiol, 37(5), 1158-1160. https://doi.org/10.1093/ije/dyn204

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Stat Med, 21(11), 1539-1558. https://doi.org/10.1002/sim.1186

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. BMJ, 327(7414), 557-560. https://doi.org/10.1136/bmj.327.7414.557

Higgins, J. P. T., & Thomas, J. (2019). Cochrane Handbook for Systematic Reviews of Interventions (J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch, Eds. 2nd Edition. ed.). Wiley.

IntHout, J., Ioannidis, J. P. A., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open, 6(7), e010247. https://doi.org/10.1136/bmjopen-2015-010247

Rice, K., Higgins, J. P. T., & Lumley, T. (2017). A re-evaluation of fixed effect(s) meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society), n/a-n/a. https://doi.org/10.1111/rssa.12275