Meta-analysis to study the promotion effect of protective factors in mental health

Liberato Camilleri, Emma Scicluna, Joeline Camilleri
Department of Statistics and Operations Research
University of Malta
Msida (MSD 06) Malta
E-mail: liberato.camilleri@um.edu.mt

KEYWORDS
Meta-analysis, Restricted maximum likelihood estimation, Egger's test, Cohens' and Hedges' effect size estimates, Forest plots, Galbraith plots, Funnel plots.

ABSTRACT

Standards for accurate and trustworthy reviews, integration and syntheses of studies that address similar research concerns are critical within the scientific world. Very often, researchers are confronted with massive amounts of results that do not always provide them with conclusive answers. By summarizing and synthesizing results of smaller studies that have inconclusive outcomes, it is more likely to produce more robust and accurate estimates and stronger inferences. In the application, this paper combines the results of six studies carried out in six European countries to investigate the promotion effect of protective factors in mental health, namely social emotional, learning, resilience and prosocial behaviour.

1. Introduction

The term 'meta-analysis' was coined in 1976 by the statistician Gene Glass to refer to the integration of research findings from a collection of analytic results elicited from individual studies. Meta-analysis grew out of necessity when the amount of research studies in various fields was growing so fast that researches became conscious that it was necessary to come up with a technique that combines various study results. Thus, a number of researchers began to develop a number of distinct methods in order to combine these related study results.

The birth of meta-analysis was marked when Glass delivered a speech on his finding of a new quantitative scientific analysis. His seminal contribution was considered as a breakthrough in all sciences and since then, meta-analysis has grown at a rapid pace, and its application has spread from one discipline to another. Glass used standard statistical procedures, mainly analysis of variance and regression analysis, but he applied these techniques to summary statistics, rather than to raw data. Other meta-analystscontributed to develop their own statistical methods to improve and simplify the statistical procedures. The work of Glass was extended by John Hunter and Frank Schmidt who developed a set of general procedures for test validity studies. Their research on validity generalization shared many similarities with Glass' research on meta-analytic methodology.

Larry Hedges demonstrated that the effect size statistics usually calculated for meta-analyses were biased estimators of an underlying population effect. In addition, Hedges developed an accurate estimate of the standard error for operative effect sizes. Robert Rosenthal compared and combined the findings of 311 studies that focussed on the experimeters' biasing effect on their results.Another contribution included the measurement of the size of study effects and the standardized mean difference between experimental and control groups, as well as statistical analyses of the relationship between study features and the standardized mean difference between an experimental and a control group. Moreover, Rosenthal proposed the use of contrast weights in studies known as focused statistical tests. Rosenthal used these focused tests to determine whether certain studies have a greater impact than others.

Since 1976, the meta-analytic methodology has continued to grow. Several publications outlining meta-analysis approaches were published in the early 1980's. Some publications include those of Glass, McGraw and Smith in 1981; Hunter, Schmidt, and Jackson in 1982; and Rosenthal in 1984. In 1985, Hedges and Olkin published a book on meta-analysis based on classical statistics. This was crucial in instilling formality and statistical validity in the approach, as well as serving as a springboard for further meta-analytic developments. In the last 25 years, meta-analysis has evolved from a small number of contributions by statisticians to synthesize study findings to a comprehensive statistical technique that provides more precise estimates of effect sizes and increase the generalisabity of individual study results. Its significance in the social sciences and education is negligible in comparison to its impact in medicine, where hundreds of meta-analyses have been published. All in all, meta-analysis can be seen as the way of the future when it comes to synthesizing study findings.

2. Effect Sizes based on Means

The between-group mean difference is a common effect size used in meta-analysis, which can only be used when all related studies use the same scale to measure and evaluate the outcome of interest. This allows the meta-analysis to be performed on the raw difference in means between two independent groups. Consider two normal populations having means μ_1 and μ_2 and standard deviations σ_1 and σ_2. Suppose that two samples of sizes n_1 and n_2 are selected from these two populations, each having means $\bar{X}_1$ and $\bar{X}_2$ and standard deviations S_1 and S_2. The sample mean difference $D = \bar{X}_1 - \bar{X}_2$ is the effect size estimate of $\mu_1 - \mu_2$. If the two population standard deviations are equal to σ, the variance of the sampling distribution of D is given by:

$$\text{var}(D) = S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right) \text{ where } S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

If the two population standard deviations are unequal, the variance of the sampling distribution of D is given by:

$$\text{var}(D) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

The between-group standardized mean difference is an alternative effect size. It does not make sense to combine raw mean differences if various studies employ different scales to assess an outcome. In this case, the mean difference of each study is divided by the standard deviation of the study to produce an index that can be compared across all the different studies. The population standardized mean δ is defined as:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

The standardized mean difference has two versions of the effect size estimate, Cohens'*d* and Hedges' *g*. The sample estimate of the standardized mean difference drawn from two independent groups is given by:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S_p} \text{ where } S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

The variance of *d* is given by:

$$\text{var}(d) = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$$

However, d has a bias and tends to overestimate δ in small sample sizes. To eliminate this bias, *d* is converted to *g*, which is an unbiased estimate of δ, using a correction factor *J*.

$$J = 1 - \frac{3}{4df - 1} \text{ where } df = n_1 + n_2 - 1$$

The equation to convert Cohens' *d* to Hedges' *g* is given by:

$$g = Jd \text{ where } \text{var}(g) = J^2 \text{var}(d)$$

In general, *J* is always less than 1 and so *g* will always be less than *d* in absolute value; while the variance of *g* will always be less than the variance of *d*.

3. Statistical Models for Meta-Analysis

The two most popular statistical models for meta-analysis include the fixed effect model and the random effects model. The statistical methods used to combine study results when fixed effects differ from the methods used with random effects. The majority of meta-analyses assign weights to each study based on the inverse of the overall study error variance. Studies with a more precise estimate of the population effect size are given more weight than studies with a less accurate estimate. This method is applicable to both fixed effects and random effects models.

In a fixed effect model, all the studies in the meta-analysis share a common true effect size. The observed effect size deviates from the true effect size only due to a sampling error. In the fixed effects model, the common effect size is denoted by θ and the deviation of the study estimate $\hat{\theta}_i$ from θ is ε_i. The fixed effect model can therefore be formulated as follows:

$$\hat{\theta}_i = \theta + \varepsilon_i \text{ where } \varepsilon_i \sim N(0,\ \sigma_i^2)$$

where σ_i^2 is the within-study variance for study *i*. The weight w_i assigned to each individual study in a fixed effects model is the reciprocal of σ_i^2 and $\hat{\theta}$ is the weighted average of $\hat{\theta}_i$:

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i} \text{ where } w_i = \frac{1}{\sigma_i^2}$$

Using the central limit theorem, the 95% confidence interval for the summary effect θ is given by:

$$\hat{\theta} \pm 1.96\sqrt{\text{var}(\hat{\theta})} \text{ where } \text{var}(\hat{\theta}) = 1 / \sum_{i=1}^{k} w_i$$

The assumption that the true effect is the same across all studies may not be realistic. Rather than assuming that there is one true effect, one can assume that there is a distribution of true effect sizes. This gives rise to the random effects model:

$$\hat{\theta}_i = \theta_i + \epsilon_i = \mu + \zeta_i + \epsilon_i$$

where μ is the mean of the distribution of effects, ζ_i is the between-study error and ϵ_i is the sampling error associated with study *i*. Under the random effects model, one needs to take account of two levels of sampling and two sources of error; the between-study error $\zeta_i \sim N(0,\ \tau^2)$ and the sampling error $\epsilon_i \sim N(0,\ \sigma_i^2)$. Thus, in assigning weights to estimate μ both the within-studies sampling error and between studies sampling error have to be taken into account.

Similar to the fixed effects model, the weight w_i^* assigned to each study will be the inverse of its variance. However, this time the variance includes the within-studies variance σ_i^2 plus the between-studies variance τ^2. Since τ^2 is unknown it has to be estimated. There are several methods to estimate τ^2. The most common method to estimate τ^2 is the restricted maximum likelihood (REML) method. The weighted mean $\hat{\theta}^*$ is:

$$\hat{\theta}^* = \frac{\sum_{i=1}^{k} w_i^* \hat{\theta}_i}{\sum_{i=1}^{k} w_i^*} \text{ where } w_i^* = \frac{1}{\sigma_i^2 + \tau^2}$$

Using the central limit theorem, the 95% confidence interval for the summary effect θ^* is given by:

$$\hat{\theta}^* \pm 1.96\sqrt{\text{var}(\hat{\theta}^*)} \text{ where } \text{var}(\hat{\theta}^*) = 1 / \sum_{i=1}^{k} w_i^*$$

In general, the variance, standard error and confidence interval for the summary effect will always be larger under the random effects model than under the fixed effect model.

4. Graphical Displays used in Meta-Analysis

Forest plots are the most common displays to visualize meta-analysis results. Besides the observed effect size, these plots, display the confidence interval and the weight of each study. They also display the pooled effect that is calculated in the meta-analysis. Moreover, the plot makes it quite easy to identify potential outliers. Overall, a forest plot takes all relevant studies related to the same research question and allows the analyst to examine the precision of the studies and how the pooled effect relates to the observed effect sizes.

Funnel plots are scatterplot used in meta analysis that display the standard error against the effect size estimate (Hedges's g). It is used mainly as a visual aid for detecting bias or systematic heterogeneity. A symmetric funnel shape arises from a well-behaved data set, in which publication bias is unlikely. An asymmetric funnel indicates the possibility of either publication bias or a systematic difference between studies of higher and lower precision.

Galbraith plots are graphical displays that provide information about the effect sizes and their precision along with the overall effect size and potential outliers. Moreover, these plots also assist in assessing heterogeneity through the standardised effect estimate. Standardised effect estmates that are considerably distant from 0 corresponds to studies that are less robust and precise. In most Galbraith plots, the standardized effect size is plotted against the inverse of the standard error (precision). The line through the origin is the unweighted regression line, while the two parallel lines represent the 95% confidence interval.

5. Tests to identify Publication Bias

The Egger's regression test to evaluate funnel plot asymmetry is preferred to its visual representation as subjectivity is eliminated by means of statistically significant results. Egger's test regresses the standardized effect sizes on their precision and is based on a simple linear regression model which tests $H_0 : \beta_0 = 0$ against $H_1 : \beta_0 \neq 0$. Egger's regression model is:

$$\frac{\hat{\theta}_i}{s.e(\hat{\theta}_i)} = \beta_0 + \frac{\beta_1}{s.e(\hat{\theta}_i)} + \varepsilon_i \text{ where } \varepsilon_i \sim N\left(0, \sigma^2\right)$$

When no asymmetry is present in the funnel plot, the intercept of the regression model is zero. However, when it differs from zero indicates asymmetry in the funnel plot and the possibility of publication bias. Most meta-analysts report the *p*-value of Egger's regression test rather than the size of the intercept. This is due to to the fact that the magnitude of the interept provides little insight of the publication bias, unless compared with the standard error of the intercept. So, if the *p*-value of Egger's regression test is less than the 0.05 level of significance, then the null hypothesis is rejected concluding that the funnel plot is in fact not symmetric due to the possibility of publication bias.
The PET-PEESE is a modification to Egger's regression test and is a combination of two methods: the precision-effect test (PET) and the precision-effect estimate with standard error (PEESE). The Trim and Fill method is used for correcting publication bias. It is an iterative procedure and recalculates the effect size at each iterations until the plot is symmetric.

6. Tests to identify Heterogeneity

The random effects model has two sources of variation, the between-study error and the sampling error. These both cause the observed effects to differ between studies. To estimate the between-study variance τ^2, the restricted maximum likelihood estimator (REML) is used and is given by:

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{k}[w_i^*(\hat{\theta}_i - \hat{\theta}^*)]^2 - \sigma_i^2}{\sum_{i=1}^{k}(w_i^*)^2} + \frac{1}{\sum_{i=1}^{k} w_i^*}$$

where *k* is the number of studies, w_i^* are the weights, σ_i^2 are the within-study variances, $\hat{\theta}_i$ are the estimated effect sizes and $\hat{\theta}^*$ is their weighted mean. To determine whether τ^2 is significant, the Cochran's Q test is used to test $H_0 : \tau^2 = 0$ against $H_1 : \tau^2 > 0$, where the *Q* statistic is given by:

$$Q = \sum_{i=1}^{k} w_i^* \hat{\theta}_i - \frac{\left(\sum_{i=1}^{k} w_i^* \hat{\theta}_i\right)^2}{\sum_{i=1}^{k} w_i^*}$$

Q has a chi square distribution with *k*-1 degrees of freedom.

The I^2 statistic is an alternative way to quantify and report between-study heterogeneity in a meta-analysis. This statistic measures the extent of heterogeneity rather than stating whether it is present or not. It is defined as the percentage of variability in the effect sizes that is not caused by sampling error. It quantifies, in percent, how much the observed value of *Q* exceeds the expected *Q* value when there is no heterogeneity.

$$I^2 = \frac{Q - df}{Q} \text{ x } 100\% \text{ where } df = k - 1$$

The H^2 statistic is similar to the I^2 statistic and is also derived from Cochran's *Q*. It describes the ratio of the observed variation which is measured by *Q* and the expected variance due to sampling error:

$$H^2 = \frac{Q}{k-1}$$

$I^2 = 0$ and $H^2 = 1$ imply no between-study heterogeneity.

7. Application

The main objective of this application is to synthesize the results of studies carried out on adolescents from six different countries (Croatia, Greece, Italy, Latvia, Portugal, Romania)) on three latent variables, including social emotional learning (SEL), resilience and prosocial behaviour. The studies compare mean scores on these three latent variables between treatment groups, who received promotion of protective factors in mental health, and control groups who received no promotion.

The prosocial scale was generated by averaging the rating scores provided to five items taken from the SDQ questionnaire (Goodman 1997). The scale ranges from 1 to 3, where the larger the mean score the higher is the intention to help others.

The social emotional learning scale was generated by averaging the rating scores provided to twenty items taken from the SSIS-SEL questionnaire (Elliot et al. 2020). The scale ranges from 1 to 4, where the larger the mean score the higher is the level of self-awareness, relationship skills, responsible decision making, self-management, and social awareness. The resilience scale was generated by averaging the rating scores provided to ten items taken from the CD-RISC questionnaire (Connor et al. 2003). The scale ranges from 1 to 5, where the larger the mean score the higher is the student's capacity to recover quickly from difficulties.

	Treatment			Control		
Study	N	Mean	SD	N	Mean	SD
Croatia	57	3.33	.302	51	3.26	.326
Greece	42	3.09	.382	47	3.07	.334
Italy	193	3.08	.412	148	2.96	.333
Latvia	340	3.07	.361	353	3	.392
Portugal	336	3.22	.381	227	3.2	.383
Romania	54	3.54	.238	55	3.33	.374
Heterogeneity: τ^2 = 0.02, I^2 = 50.94%, H^2 = 2.04						
Test of $\theta_i = \theta_j$: Q(5) = 10.19, p = 0.07						

Table 1: Meta-analysis summary on SEL

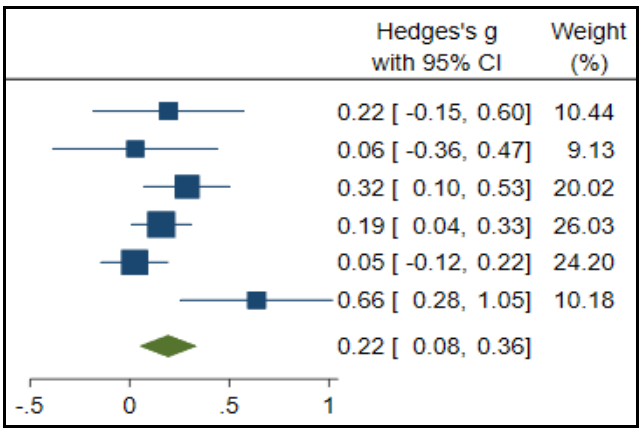

Figure 1: Forest plot displaying effect size estimates for SEL

The presence of heterogeneity can also be deduced from the heterogeneity statistics displayed in Table 1. The estimate of the between-study variance of the true effect size τ^2 is 0.02. The I^2 statistic (50.94%) indicates that around 51% of the variability in the effect size estimate is a result of between-study differences. The H^2 statistic (2.04) is larger than 1 indicating the presence of between-study heterogeneity. Moreover, the heterogeneity test yields a p-value (0.07) which is slightly larger than the 0.05 level of significance, indicating some heterogeneity between the studies exists but is not significant.

The forest plot displayed in Figure 1 shows that, for the studies combined, the difference in mean SEL scores between the treatment and control groups is 0.22. Moreover, this difference is significant since the 95% confidence interval of [0.08, 0.36]. excludes 0. The Galbraith and funnel plots, displayed in Figures 2 and 3 respectively, show that heterogeneity and assymmetry are caused mainly by the Romanian study.

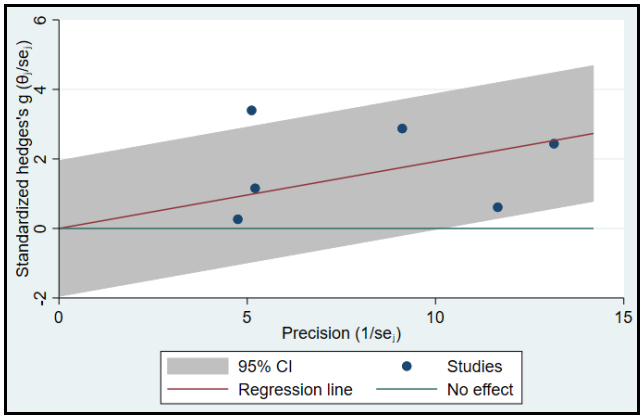

Figure 2: Galbraith plot assessing heterogeneity for SEL

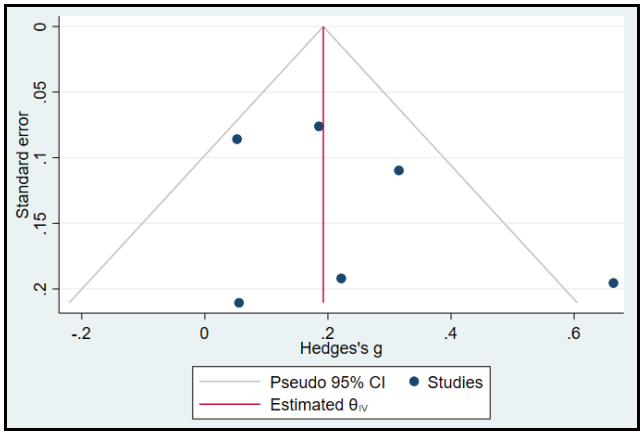

Figure 3: Funnel plot assessing publication bias for SEL

	Treatment			Control		
Study	N	Mean	SD	N	Mean	SD
Croatia	57	3.83	.54	51	3.77	.567
Greece	42	3.6	.648	47	3.57	.627
Italy	193	3.15	.766	148	3.2	.822
Latvia	340	3.58	.68	353	3.5	.693
Portugal	336	3.47	.733	227	3.48	.777
Romania	54	4.01	.581	55	3.77	.714
Heterogeneity: τ^2 = 0.00, I^2 = 4.06%, H^2 = 1.04						
Test of $\theta_i = \theta_j$: Q(5) = 5.16, p = 0.40						

Table 2: Meta-analysis summary on SEL

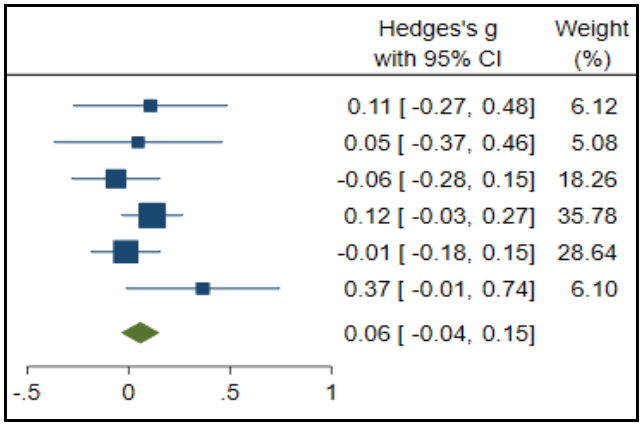

Figure 4: Forest plot displaying effect estimates for Resilience

The heterogeneity test in Table 2 yields a p-value (0.40) which is larger than the 0.05 level of significance, indicating negligible heterogeneity between the studies. The forest plot displayed in Figure 4 shows that, for the studies combined, the difference in mean resilience scores between the treatment and control groups is 0.06. Howver, this difference is not significant since the 95% confidence interval of [-0.04, 0.15]. includes 0. The Galbraith and funnel plots, displayed in Figures 5 and 6 respectively, show no heterogeneity and assymmetry between the studies.

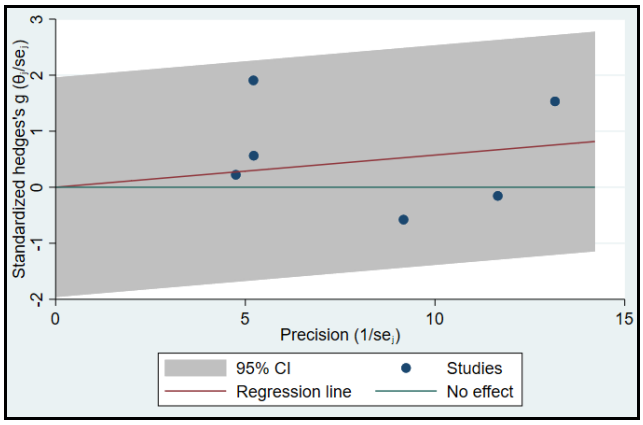

Figure 5: Galbraith plot assessing heterogeneity for Resilience

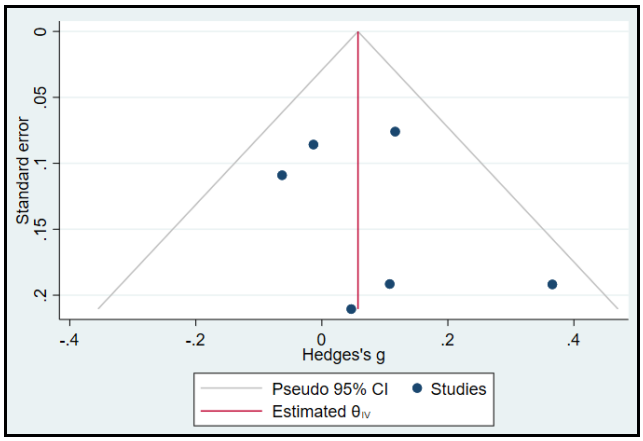

Figure 6: Funnel plot assessing publication bias for Resilience

	Treatment			Control		
Study	N	Mean	SD	N	Mean	SD
Croatia	57	2.68	.327	51	2.6	.353
Greece	42	2.62	.346	47	2.6	.27
Italy	193	2.55	.385	148	2.52	.343
Latvia	340	2.42	.35	353	2.39	.389
Portugal	336	2.63	.374	227	2.61	.349
Romania	54	2.81	.19	55	2.67	.342
Heterogeneity: τ^2 = 0.00, I^2 = 0.00%, H^2 = 1.00						
Test of $\theta_i = \theta_j$: Q(5) = 5.18, p = 0.39						

Table 3: Meta-analysis summary on Prosocial

The heterogeneity test in Table 3 yields a p-value (0.39) which is larger than the 0.05 level of significance, indicating minor heterogeneity between the studies. The forest plot displayed in Figure 7 shows that, for the studies combined, the difference in mean prosocial scores between the treatment and control groups is 0.11. Moreover, this difference is significant since the 95% confidence interval of [0.02, 0.20]. excludes 0. The Galbraith and funnel plots, displayed in Figures 8 and 9 respectively, show that the minor heterogeneity and assymmetry are caused mainly by the Romanian study.

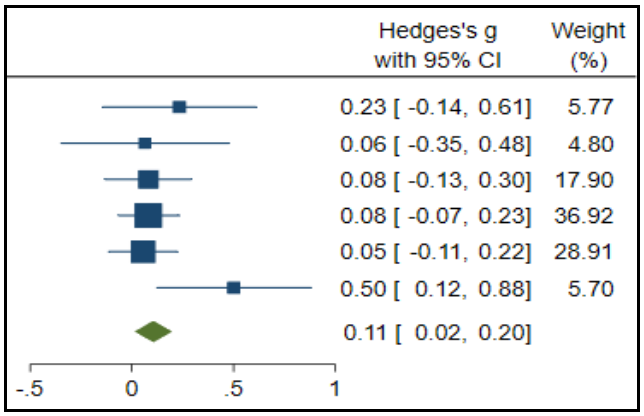

Figure 7: Forest plot displaying effect estimates for Prosocial

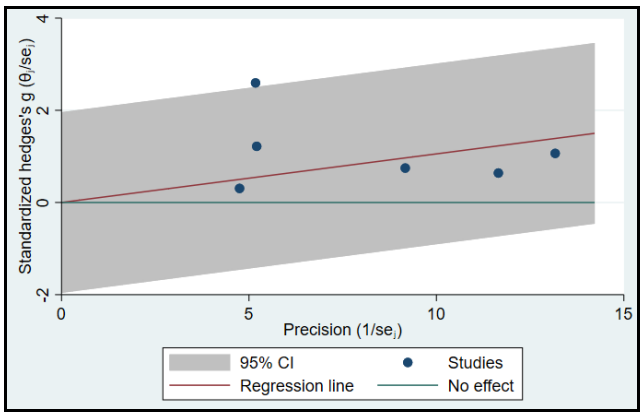

Figure 8: Galbraith plot assessing heterogeneity for Prosocial

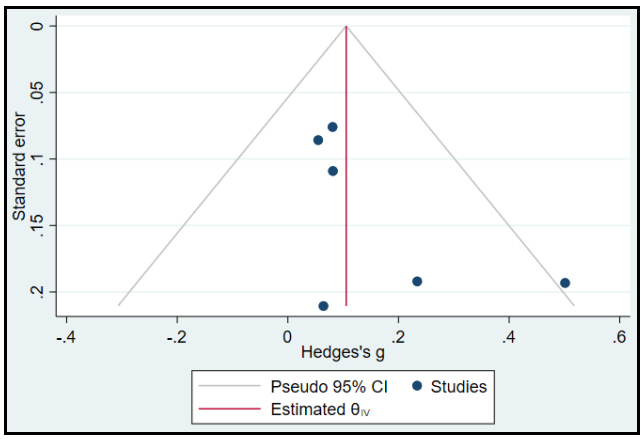

Figure 9: Funnel plot assessing publication bias for Prosocial

8. Conclusion

The forest plots show that the majority of the small studies carried out by each country yielded insignificant improvement when comparing the treatment with the control group in SEL, resilience and prosocial behaviour. However, the improvements become significant when the studies are combined. Hence we can generalize that promotion of protective factors in mental health enhances SEL and prosocial behaviour in adolescents.

9. References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cooper, H. (1998). *Integrating research: A guide for literature reviews* (3rd ed.). Newbury Park, CA: Sage.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Connor, K. M., and Davidson, J. R. (2003). Development of a new resilience scale: The Connor-Davidson resilience scale (CD-RISC). *Depress. Anxiety* 18:2, 76-82.

Egger, M., Smith, G. D., & Altman, D. G. (Eds.). (2001*). Systematic reviews in healthcare: Meta-analysis in context* (2nd ed.). London: BMJ Publishing Group.

Elliott, S. N., DiPerna, J. C., Anthony, C. J., Lei, P., and Gresham, F. M. (2020). *Social Skills Improvement System, Social Emotional Learning (SSIS-SEL) Brief Scales*. Scottsdale, AZ: SAIL CoLab.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *J. Child Psychol. Psychiatry* 38:5, 581-586.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *EducationalResearch*, *5*, 3-8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Newbury Park, CA: Sage.

Hardy, R. J., & Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine, 15*, 619-629.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *Thehandbook of research synthesis* (pp. 285-299). New York: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper and L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-321). New York: Russell Sage Foundation.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev. ed.). Newbury Park, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper and L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research*. Chichester, UK: Wiley.