View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Erasmus University Digital Repository

How to interpret results of meta-analysis

Tony Hak, Henk van Rhee, & Robert Suurmond

Version 1.0, March 2016

Meta-analysis is a systematic method for synthesizing quantitative results of different empirical studies regarding the effect of an independent variable (or determinant, or intervention, or treatment) on a defined outcome (or dependent variable). Mainly developed in medical and psychological research as a tool for synthesizing empirical information about the outcomes of a treatment, meta-analysis is now increasingly used in the social sciences as a tool for hypothesis testing. However, the assumptions underlying meta-analytic hypothesis testing in the social sciences will usually not be met under real-life conditions. This is the reason why meta-analysis is increasingly conducted with a different aim, based on more realistic assumptions. That aim is to explore the dispersion of effect sizes.

Preferred citation of this text:

Hak, T., Van Rhee, H. J., & Suurmond, R. (2016). How to interpret results of meta-analysis. (Version 1.0) Rotterdam, The Netherlands: Erasmus Rotterdam Institute of Management. www.erim.eur.nl/research-support/meta-essentials/downloads

The *Meta-Essentials* workbooks are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/.

Contact

- Tony Hak, thak@rsm.nl
  Rotterdam School of Management
  Burgemeester Oudlaan 50
  3062PA Rotterdam, The Netherlands
- Henk van Rhee, vanrhee@rsm.nl
- Robert Suurmond, suurmond@rsm.nl

Contents

1 Introduction

Meta-analysis is a systematic method for synthesizing quantitative results of different empirical studies regarding the effect of an independent variable (or determinant, or intervention, or treatment) on a defined outcome (or dependent variable). Mainly developed in medical and psychological research as a tool for synthesizing empirical information about the outcomes of a treatment, meta-analysis is now increasingly used in the social sciences as a tool for *hypothesis testing*. However, the assumptions underlying meta-analytic hypothesis testing in the social sciences will usually not be met under real-life conditions. This is the reason why meta-analysis is increasingly conducted with a different aim, based on more realistic assumptions. That aim is to explore the *dispersion of effect sizes*.

The structure of this document follows the structure of the workbooks of *Meta-Essentials*.

Assuming that the appropriate workbook has been chosen and that the relevant information about the different studies has been entered on the Input sheet of that workbook, it discusses the interpretation of the forest plot, subgroup analysis, moderator analysis, and publication bias analyses.

The aim of this document is to support the researcher in interpreting the results of a meta-analysis. Throughout this text we will use statistics, figures, and tables as provided by *Meta-Essentials*, user-friendly software for meta-analysis that is freely downloadable (http://www.erim.eur.nl/research-support/meta-essentials/). Although the figures and tables in this document are taken from examples in *Meta-Essentials*, its contents are applicable to any meta-analysis irrespective of the software package that is used.

When the *Meta-Essentials* software is downloaded from the website, its workbooks are already filled with (fictional) data. The examples and screen-prints that are used in the user manual are taken from Workbook 1 (*Effect size data.xls*). The same examples and screen-prints are used in this document. This means that the reader can generate and manipulate these same examples in that workbook, which might be useful for a critical reading of the following text.

1.1 Assumptions

In every meta-analysis the following assumptions should be made, and the researcher is supposed to have verified that they are true for the meta-analysis at hand:

1. An *effect* is precisely defined, i.e., an *independent* as well as a *dependent* variable are defined, and all studies in the meta-analysis are empirical studies of that effect. These definitions should be precise enough to allow the researcher to include (and exclude) studies on transparent grounds.

   Note: This might seem to be an obvious assumption, but it occurs quite often that authors claim that they have studied an effect of some independent variable on a dependent variable whereas on closer inspection it appears that they have studied other variables and, hence, another effect.
2. The type of *unit or object* in which this effect might occur is specified (e.g., persons, countries, teams, specific types of organisational units) and the *domain* for which the effect will be meta-analysed is clearly delimited (e.g., all persons, not all persons but only adults, or only women; all countries, not all countries but only developed countries; all teams, not all teams but only product development teams in specific industries; all marketing departments, not all marketing departments but only marketing departments in a specific economic sector).
3. Assuming that the researcher's aim is to synthesize empirical results about the effect in a domain (e.g., all patients in the world who might benefit from a specific treatment), *all* empirical studies of the effect in that domain should have been identified.

   Note: This is a problematic assumption. Usually the set of studies that is meta-analysed is not complete because some studies have not been published, or have been published in a form to which the researcher has no access, or have been published in a language that the researcher cannot read, etcetera.
4. All studies are *methodologically sound*, i.e., data have been collected from a complete probability sample of a defined population, measurement has been valid and reliable, and the statistical analysis has been adequate.

   Note: This is also a problematic assumption because most studies fail one or more of these criteria: the population might not have been specified, probability sampling might not have been conducted, there might be missing cases, measurement might not be valid or reliable, statistical procedures might be inappropriate (e.g., when statistical methods for differences between independent groups are used in a pretest - posttest design). Note that verification of this assumption requires a methodological evaluation of each study, irrespective of its source or reputation ("peer-reviewed", "highly cited", "good journal", etc.). If this quality requirement is neglected or violated, then any meta-analytic result is meaningless (garbage in, garbage out).
5. *Effect size measures* in these studies are comparable. Specifically, they need to have the same scale across studies.

Assumptions 1, 2, 4 and 5 will not be further discussed in this document. In other words, it is assumed throughout this text that the researcher has, as an input for meta-analysis, a set of *comparable and methodologically sound effect sizes for specified populations from a domain*. Assumption 3 will be further discussed on several occasions below.

2 The forest plot

The main outcome of any meta-analysis is a forest plot, a graphical display as in Figure 1, which is an example of a forest plot generated with Workbook 1 (*Effect size data.xls*) of *Meta-Essentials*. This is the same plot as is used as an example in the User Manual. The x-axis forms the effect size scale, plotted on the top of the plot. Each row, except the bottom one, represents a study's effect size estimate in the form of a point and a (95%) confidence interval. This is the statistically correct way of representing the results of a single study, namely as an estimate of an interval in which the "true" effect (in the population) will most probably lie. Remember that it is assumed that every study in the meta-analysis is a study of a complete probability sample of a specified population. If this assumption is not met in a study, no inference can be made from the "sample" to a population and hence, comparing the observed effect size with observations in other studies is not meaningful.

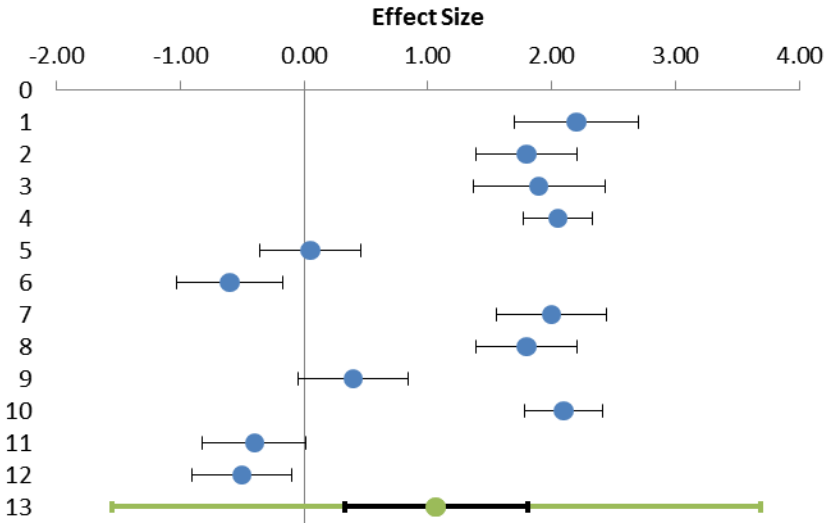


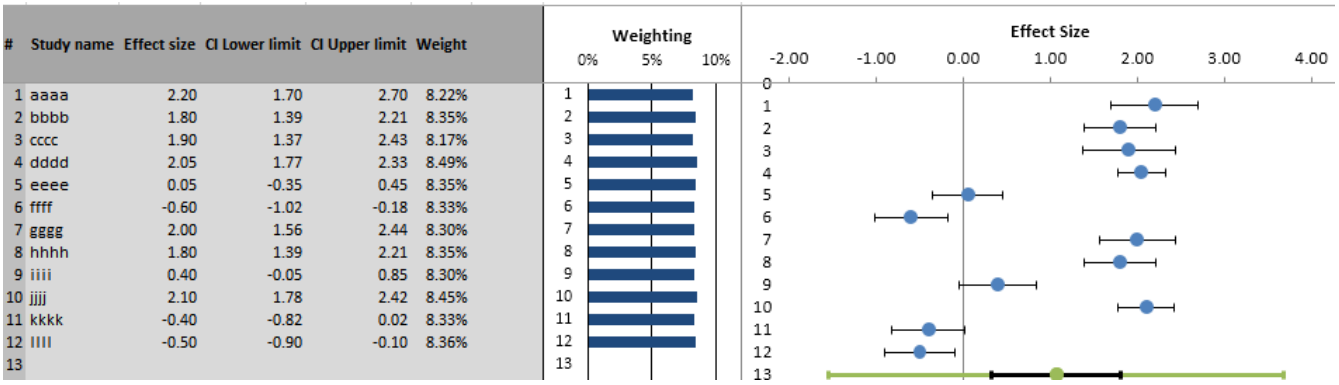
Figure 1: Example of a forest plot in Meta-Essentials

The point estimate is represented in the forest plot by a smaller or a larger bullet. The relative size of these bullets represents a study's weight in the generation of the meta-analytic result

The plot presented in Figure 1 is fictitious and constructed for illustration purposes. Its characteristics are typical for forest plots in the social sciences:

- Some confidence intervals are entirely on the positive side of zero. In traditional terminology, these studies show a statistically significant positive effect.
- Other confidence intervals are entirely on the negative side of zero. In traditional terminology, these studies show a statistically significant negative effect.
- Other confidence intervals include zero. In traditional terminology, these studies show an effect that is not statistically significant.

The forest plot, discussed so far, is just a pictorial representation of results of a set of studies. The same information (point estimates with confidence intervals, and weights, for every study) could also have been expressed by numbers in a table. This table is actually also available on the forest plot sheet in *Meta-Essentials*, on the left side of the plot (see Figure 2).



#	Study name	Effect size	CI Lower limit	CI Upper limit	Weight
1	aaaa	2.20	1.70	2.70	8.22%
2	bbbb	1.80	1.39	2.21	8.35%
3	cccc	1.90	1.37	2.43	8.17%
4	dddd	2.05	1.77	2.33	8.49%
5	eeee	0.05	-0.35	0.45	8.35%
6	ffff	-0.60	-1.02	-0.18	8.33%
7	gggg	2.00	1.56	2.44	8.30%
8	hhhh	1.80	1.39	2.21	8.35%
9	iiii	0.40	-0.05	0.85	8.30%
10	jjjj	2.10	1.78	2.42	8.45%
11	kkkk	-0.40	-0.82	0.02	8.33%
12	llll	-0.50	-0.90	-0.10	8.36%
13					

Figure 2: Part of forest plot sheet in Meta-Essentials, with a table and its corresponding pictorial representation

The bottom row (or "summary row") of the forest plot turns the plot into a "meta-analysis". This is the row that represents the result of the meta-analysis. In *Meta-Essentials* this meta-analytic result (line 13 in Figure 1) consists of two intervals, both around the same bullet. This bullet represents the *weighted average effect*, to which we will refer as "combined" effect size. The smaller, black, interval is a *confidence interval*. The larger, green, interval is the *prediction interval*. We will discuss the interpretation of both intervals in the subsequent sections.

2.1 Confidence interval: hypothesis testing

The confidence interval of the combined effect size in Figure 1 does not include zero, i.e., in case of a confidence level of 95% the *p*-value is smaller than .05. In traditional terminology, this means that the meta-analytic effect is statistically significant. If the aim of the meta-analysis is to test the hypothesis that there is an effect, then the null hypothesis can be rejected and the alternative hypothesis (that there is an effect) is deemed more likely in this example. The user can find the corresponding Z-value and *p*-values (one-tailed and two-tailed) in the column on the left side of the forest plot sheet in *Meta-Essentials* (see Figure 3).

Z-value	3.18
One-tailed p-value	0.001
Two-tailed p-value	0.001

Figure 3: Part of forest plot sheet in Meta-Essentials, with Z-value and p-values.

The forest plot in Figure 1 shows that research results have been "contradictory" or "ambiguous". Some studies have shown statistically significant positive effects. Other studies have shown statistically significant negative effects. And there have also been some studies with effects that are statistically non-significant. In standard practice, meta-analysis is aimed at "solving" the problem that results of studies differ and apparently contradict each other, by generating a "*combined*" effect. The combined effect might be significantly different from zero (or not). This type of result of a standard practice meta-analysis is considered very valuable because it clearly is a solution to the problem of "contradictory" evidence or, more often, to the problem of "insignificant results". Because meta-analysis functions as a more powerful significance test it generates a more useful and more convincing result than a single study. For example, suppose that the forest plot in Figure 1 is a plot of the results of studies of the effects of a specific educational reform intervention. These studies have been conducted in different countries, in different contexts (e.g., in schools managed by the government and schools managed by non-governmental organisations), and with different types of students (e.g., different genders, different ages, different social backgrounds). The statistical significance of the combined effect might now be interpreted as evidence (or "statistical proof") that this specific intervention "significantly" improves educational outcomes. This result, then, might be the basis of a policy decision to invest in this type of educational intervention.

However, a Z-value or *p*-value is not an effect size. A government or an education minister is not merely interested in the fact that there is a positive effect of an intervention but also in how large the effect is. Statistical significance loses its relevance when "samples" get very large. Because the pooled sample size in a meta-analysis is usually very large, the combined effect will almost certainly be significant, even if the combined effect size is very small. If not, a statistically significant combined effect can be generated by adding studies to the meta-analysis. If the combined effect is taken as the basis for a decision about implementing this intervention, its estimated effect size and its precision (or lack of it as indicated by the width of the confidence interval) should be interpreted. In this example, policy makers should decide

whether an effect size of 0.20 (the lower bound of the confidence interval of the estimate of the combined effect) is large enough for deciding in favor of the intervention.

A serious problem with this approach is that none of the usual requirements for null hypothesis significance testing is met in a meta-analysis. The core requirement for a significance test (or for any form of inferential statistics in general) is that the effect is observed in a random (or probability) sample from a defined population (enabling a sampling frame). In meta-analysis, however, there is no population and there is no probability sample. The input of the meta-analysis is results of studies. These studies have generated estimates of effect sizes in populations, which are represented in the forest plot by point estimates and their confidence intervals. This means that the sample in a meta-analysis (if any) is a sample of populations and that the combined effect size (and its confidence interval) that is calculated in the meta-analysis is an inference about the effect in the population of these populations. This super-population might coincide with the domain of the study (as specified in assumption 3 above). However, it is obvious that the set of populations that have been studied (and of which the results have been entered in the meta-analysis) is not (and cannot be) a probability sample from that domain. Usually, a study is conducted in a sample (if a sample is drawn at all) from a population that is chosen by convenience. A sample for a study of an intervention, for instance, is normally drawn from a population to which the researcher happens to have access (e.g., the schools in a school district of which a gatekeeper happens to be a friend or former colleague of the researcher; or the schools in a district that happens to be managed by an innovative governor). Because a set of effect sizes in a meta-analysis is not a probability sample, null hypothesis significance testing is statistically not appropriate.

The fact that populations (or studies) cannot be treated as a probability sample is usually evident from the observed effect sizes themselves. In the forest plot in Figure 1, for instance, it is obvious that the difference between the results of studies 1-4 on the one hand, and of studies 5 and 6 on the other hand, cannot be explained by sampling variation. Studies 1-4 *might* have been studies of probability samples of one population, and studies 5 and 6 might have been studies of samples from one (other) population, but it is statistically not probable that these two populations were identical to each other. For instance, studies 1-4 might have been four field experiments in a small number of adjacent school districts in one country (say, the United Kingdom, coded AA on the input sheet), whereas studies 5 and 6 might have been two field experiments in another country (BB). When the pattern of the forest plot itself suggests that there are different types of population with rather different effect sizes (as in this example), then the “combined” effect size is not a useful parameter any more.

Estimating a combined effect in a subgroup might still be useful. In this (fictitious) example, it might be useful for policy makers to have an estimate of the combined effect size of an intervention in the United Kingdom (which will be close to 2.0, which might be an effect size that is large enough to recommend the intervention) and another one in another country (which will be negative and close to zero, from which might be concluded that funding of that specific intervention should not come from the education ministry in that country). This issue will be further discussed below in the section on subgroup analysis.

2.2 Estimating the extent of heterogeneity

As suggested in the previous section, the combined effect size (and its confidence interval) is not a useful outcome of the meta-analysis as presented in Figure 1. The plot in Figure 1 itself suggests that there are different effect sizes in different types of populations. In other words, the domain that is analysed in this meta-analysis must be seen as "heterogeneous". It consists of parts (or sub-domains), each with a different "true" effect size. The forest plot sheet of *Meta-Essentials* provides numerical information about the degree of heterogeneity. Four types of information about heterogeneity are provided: the *Q-statistic* with a *p*-value; I^2; T^2; and Tau (see Figure 4).

Heterogeneity	
Q	362.77
p_q	0.000
I^2	0.97
T^2	1.31
T	1.14

Figure 4: Part of forest plot sheet in Meta-Essentials, with information about heterogeneity

The *Q-statistic* (also referred to as "Cochrane's Q") is the weighted sum of squared differences between the observed effects and the weighted average effect. (See Borenstein et al., 2009: 109-113, for how *Q* is computed.) The *Q-statistic* is only a measure of variation around the average and is not yet a measure of heterogeneity. In order to compute the heterogeneity, *Q* must be compared with the variation that would be observed if all studies were studies of a probability sample from the same population. This difference is computed in a meta-analysis with two main aims: (1) a null hypothesis significance test can be performed on this difference; and (2) it is used for calculating the other measures of heterogeneity. In the example the *p*-value is 0.000. The test of the null hypothesis is subject to the same caveats as all tests of significance (see Borenstein et al., 2009: 112-113). The *p*-value is not an effect size and, hence, is not a measure of the extent of heterogeneity. A low *p*-value only indicates that there probably is some (undetermined) degree of heterogeneity. As in any significance test, non-significance of *Q* in a study cannot be used as evidence for the null, in this case for homogeneity of the domain studied. Meta-analysts, thus, should not look at the value of *Q* nor its *p*-value. Rather, they should interpret the other measures of heterogeneity.

I^2 is a measure for the *proportion* of observed variance that reflects real differences in effect size. (See Borenstein et al., 2009: 117-119, for how I^2 is computed.) It is expressed as a percentage with a range from 0 to 100 percent. It is a *relative* measure. It is not a measure of variation in terms of the scale of the effect size parameter. Hence its usefulness is limited. Borenstein et al. (2009: 119) advise to use I^2 as a criterion for a decision whether a subgroup analysis or moderator analysis (discussed below) is indicated. If I^2 is low, then there is no heterogeneity to speak of and hence nothing to be explored in a subgroup or moderator analysis. If I^2 is large, then such an analysis is likely to be worthwhile. In the example, I^2 is 97%. This very high proportion suggests that the studies in this meta-analysis cannot be considered to be studies of the same population.

Both T^2 and *Tau* are measures of the dispersion of true effect sizes between studies in terms of the scale of the effect size. (See Borenstein et al., 2009: 114-117, for how T^2 and *Tau* are estimated.) T^2 is an estimate of the variance of the true effect sizes. Or, in the words of Borenstein et al. (2009: 114): "If we had an infinitely large sample of studies, each itself infinitely large (so that the estimate in each study was the true effect) and computed the variance of these effects, this variance would be T^2". T^2 is not used itself as a measure of heterogeneity but is used in two other ways: (1) it is used to compute *Tau*; and (2) it is used to assign weights to the studies in the meta-analysis under the random-effects model (discussed below). *Tau* is an estimate of the standard deviation of the distribution of true effect sizes, under the assumption that these true effect sizes are normally distributed. *Tau* is used for computing the *prediction interval*.

Summarizing, how should this multiple information about heterogeneity be interpreted and used? We recommend to use I^2 as the main source of information about the extent of heterogeneity. As soon as I^2 is larger than an (arbitrary) proportion (say 25%), the meta-analyst should not interpret the combined effect size as meaningful and should not conduct any form of significance testing. After such a decision has been made, the meta-analyst should focus on an analysis of the dispersion of true effect sizes, and of its determinants (moderators). *Tau* is a useful first indication of the extent of this dispersion. However, the *prediction interval* is a more direct and more easily interpretable indicator.

2.3 Prediction interval

In *Meta*-Essentials, the larger, green, interval around the combined effect size on the bottom row of the forest plot is the *prediction interval*. (See Borenstein et al., 2009: 129-131, for how a *prediction interval* is estimated.) The 95% prediction interval gives the range in which the point estimate of 95% of future studies will fall, assuming that true effect sizes are normally distributed through the domain. Because the prediction interval is estimated based on the effect sizes observed in the studies that are meta-analysed, the prediction interval corresponds more or less (depending on whether sampling variation in the meta-analysed studies is large or small) with the range of effect sizes that are meta-analysed and that are represented in the forest plot. This implies that the prediction interval can only "predict" with some accuracy if no relevant selection bias exists in the set of populations that have been studied (i.e., if the populations of which the effect size estimate is included in the meta-analysis are "representative" for the domain). If selection bias exists, then effect sizes observed in future studies might occur beyond the limits of the prediction interval. Because selection bias is more likely than not (which will be argued below), it is recommended to *interpret the prediction interval as a description of the range of observed effect sizes* rather than as a prediction of the range of effect sizes that will be observed in future studies (despite its name "prediction" interval).

2.4 Model

A meta-analyst can choose between a 'fixed effects' model and a 'random effects' model. In the '*fixed* effects' model it is assumed that all differences between effect sizes observed in different studies are only due to sampling error. In other words, it is assumed that there is no "heterogeneity". In the '*random* effects' model it is assumed that there is heterogeneity. The assumptions underlying the fixed effects model are very rarely met. Furthermore, when a fixed effects model would make sense to use, i.e., when there is little variance in effect sizes, the random effects model automatically converges into a fixed effects model. Therefore, it is strongly recommended to always use the random effects model, and to interpret the heterogeneity measures (discussed above) before deciding to use the fixed effects model (if at all).

3 Subgroup analysis

As discussed above, the forest plot in Figure 1 suggests that there are two subgroups in the domain that have a different "true" effect size. If this was a study of the effect of an intervention, it could be that we are dealing here with results from studies in different countries. It is possible to explore a hypothesis about a difference in effect between subgroups. This is, however, only possible for known characteristics of the populations that were studied. These characteristics must have been coded and entered on the Input sheet of the software. Figure 5 is an example of a forest plot on the Subgroup Analysis sheet for the same studies as have been discussed above (Figure 1). It is a bit misleading in the sense that the neat separation between the two subgroups in this plot is the result of how we have set up the example. In this fictitious example, code AA is assigned to all studies with an effect size around 2.0 and code BB has been assigned to the other studies. In actual practice, it is rare to find the effects observed in subgroups so clearly different in every single study. If study 4, for instance, would have been a study of subgroup BB, then the results would have been very different for subgroup BB.

Note that the weighted average effect (0.89) in Figure 5 is different from the one in Figure 1 (1.07) and that both the confidence interval and the prediction interval are larger. These differences between the results in these two plots can be explained by the fact that the average effect and its intervals in Figure 5 are calculated from the subgroup effects (N=2) and those in Figure 1 from the effects observed in the original studies (N=14). Therefore, it is not recommended to use information about the combined effect and its intervals from the subgroup analysis!

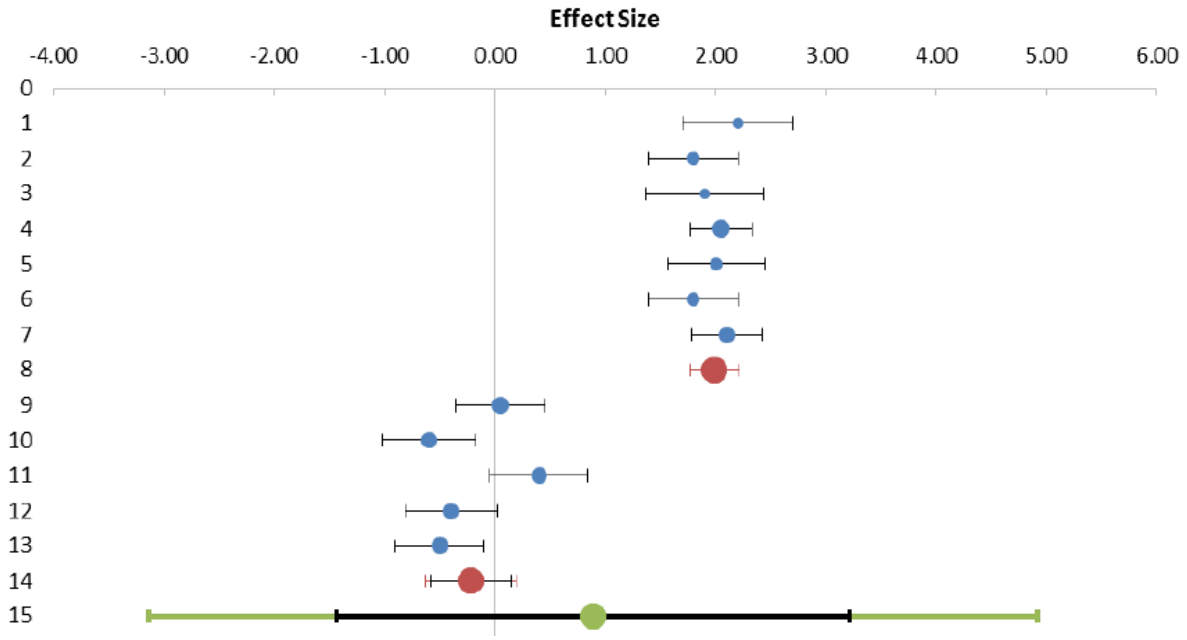


Figure 5: Example of a forest plot on the Subgroup Analysis sheet

For the interpretation of the results within a subgroup, the same principles and caveats apply as were discussed above for forest plots and meta-analytic results in general. Similarly, the meta-analyst should first interpret Information about heterogeneity. In *Meta-Essentials,* heterogeneity parameters for each subgroup are computed and presented (see Figure 6).

	#	Study name / Subgroup name	Effect size	CI LL	CI UL	Weight	Q	p_Q	I^2	T^2	T	PI LL	PI UL
	1	aaaa	2.20	1.70	2.70	9.59%							
	2	bbbb	1.80	1.39	2.21	13.36%							
	3	cccc	1.90	1.37	2.43	8.65%							
	4	dddd	2.05	1.77	2.33	23.39%							
	5	gggg	2.00	1.56	2.44	11.85%							
	6	hhhh	1.80	1.39	2.21	13.36%							
	7	jjjj	2.10	1.78	2.42	19.79%							
Subgroup AA	8	**AA**	**1.99**	**1.88**	**2.10**	**50.32%**	**3.14**	**0.791**	**0.00%**	**0.01**	**0.10**	**1.77**	**2.21**
	9	eeee	0.05	-0.35	0.45	21.07%							
	10	ffff	-0.60	-1.02	-0.18	19.50%							
	11	iiii	0.40	-0.05	0.85	18.08%							
	12	kkkk	-0.40	-0.82	0.02	19.90%							
	13	llll	-0.50	-0.90	-0.10	21.45%							
Supgroup BB	14	**BB**	**-0.22**	**-0.58**	**0.14**	**49.68%**	**15.25**	**0.004**	**73.77%**	**0.01**	**0.10**	**-0.63**	**0.19**
Combined effect size	15	**Combined effect size**	**0.89**	**-1.44**	**3.22**		**362.77**	**0.000**	**95.31%**	**1.28**	**1.13**	**-3.14**	**4.92**

Figure 6: Example of table with studies and subgroups on the Subgroup Analysis sheet

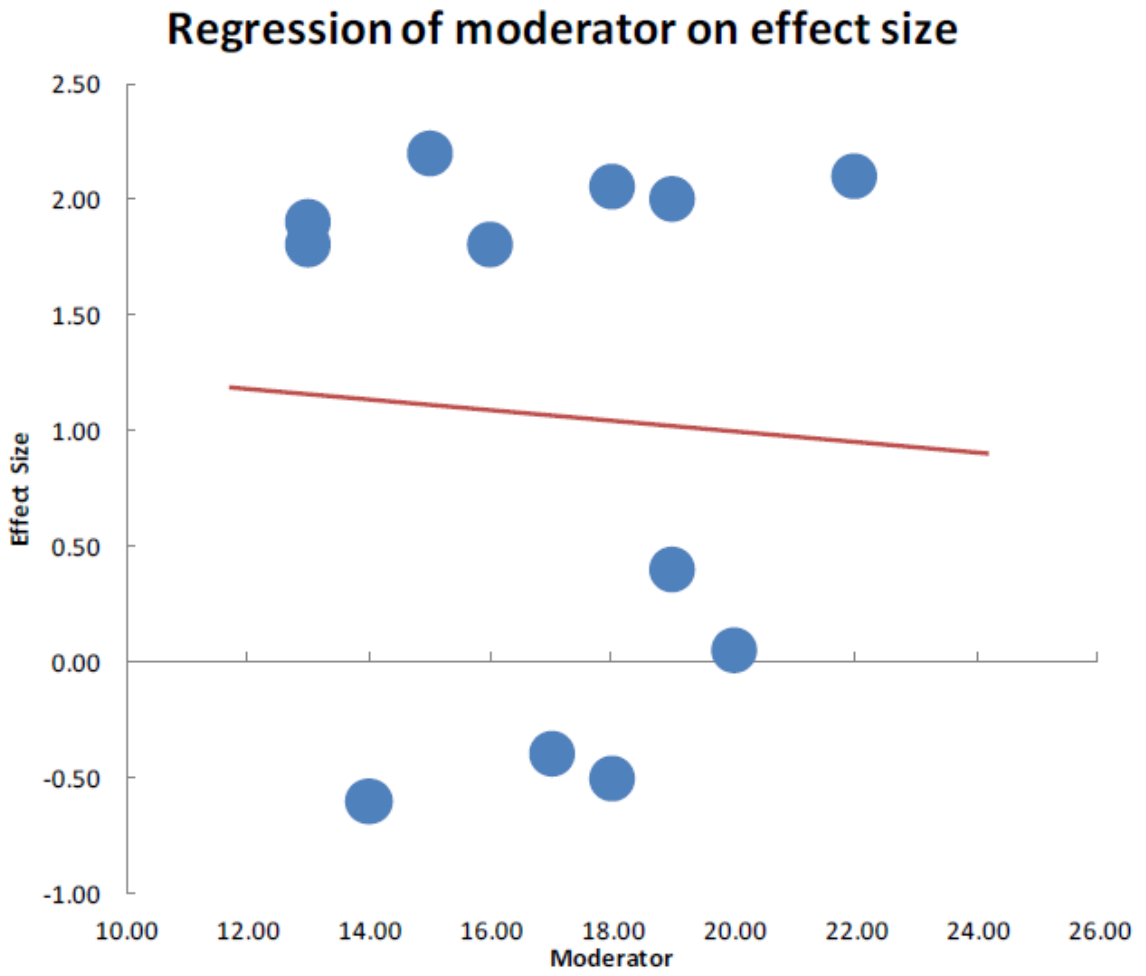
Whereas, in this example, it appeared to be erroneous to assume that the total set of studies (Figure 1) can be seen as studies of one (homogeneous) population, it appears to be possible to treat at least one of the subgroups as a homogeneous population. Figure 6 presents, among other things, the four heterogeneity parameters for each of the two subgroups. I^2 for subgroup AA is 0%, indicating that all studies in this subgroup have produced an estimate of the same "true" effect size in a homogeneous population (AA). The estimate of the effect is 1.99 (95%CI 1.88-2.10). In this subgroup, the information about the prediction interval (1.77-2.21) can be ignored. I^2 for subgroup BB is almost 74%, indicating that this subgroup is very heterogeneous and hence cannot be meta-analysed as if it is one single population. This implies that it is not useful to interpret the combined effect in subgroup BB, and hence we should focus on the prediction interval (from -0.63 to 0.19) rather than on the estimate and its confidence interval.

Note that *Meta-Essentials* does not give statistical information about the size of the difference between the average subgroup effects. Obviously, this difference can easily be calculated by the researcher from the information in Figure 6 (it is 2.21), but it is not recommended to do so for the heterogeneity of subgroup BB. There is no meaningful average effect in this subgroup and hence there is no value that can be compared with the effect in subgroup AA. The inclusion of the Subgroup analysis in *Meta-Essentials*, thus, is not primarily aimed at supporting an analysis of the differences between subgroups. It is rather aimed at supporting the identification of subgroups that might be homogeneous enough to allow estimation of a combined effect in that subgroup (as subgroup AA in the example).

If it assumed that there is no relevant selection bias in the set of studies from subgroup AA (which is a rather problematic assumption that can only be proven to be correct after inclusion of many more studies), then the average effect size might be interpreted as an estimate of the "true" effect in AA. In more practical terms, if this is a meta-analysis of the effects of an intervention on an educational outcome, then the results of the combined meta-analysis (as presented on the forest plot sheet; Figure 1) cannot be used as evidence for its effectiveness or lack of effectiveness. However, the results of the subgroup analysis (Figure 6) indicate that the treatment is effective in AA (e.g., in the United Kingdom) with an effect size around 2.0.

4 Moderator analysis

Output on the Moderator Analysis sheet consists of a scatter plot with a regression line and a table with a number of statistics (see Figure 7).



	B	SE	CI LL	CI UL	β	Z-value	p-value
Intercept	1.45	2.22	-3.43	6.33		0.66	0.512
Slope	-0.02	0.13	-0.31	0.26	-0.06	-0.18	0.860

Fit statistics	Q (Sum of squares)	df	p_Q	Mean square	F-Value	p-value
Model	0.03	1	0.860	0.03	0.03	0.863
Residual	9.92	10	0.448	0.99		
Total	9.95	11	0.535			

Combined effect size	1.07
T^2 (method of moments estimation)	1.45
R^2	0.31%

Figure 7: Example of the Moderator Analysis sheet

Although *Meta-Essentials* generates the statistics that are usually presented in a regression analysis, it is not wise to give much weight to these outcomes, particularly because there is only a small number of data points (studies). As with any regression analysis, the researcher should begin with an interpretation of the scatter plot, not of the line drawn through it. In the example in Figure 7, for instance, it is clear from the scatter plot that there is no observable relation between the moderator and the observed effect sizes. In this example, this is confirmed by the "insignificant" result of a significance test for the regression weight.

Let us assume that the moderator in this example is the average size of a class in the schools or school districts that were studied. The result of this moderator analysis, then, suggests that the class size does not have an influence on the effect of the intervention. However, these results are very sensitive for the addition of removal of a single study, and the "sample size" is so small that even a very steep regression line (or high weight) will not be significant. In Figure 8, for instance, only two studies have been removed. Visual inspection of the scatter plot after removal of two studies (Figure 8) still suggests that there is hardly any relation between the value of the moderator and the observed effect size. The regression line, however, is steep.

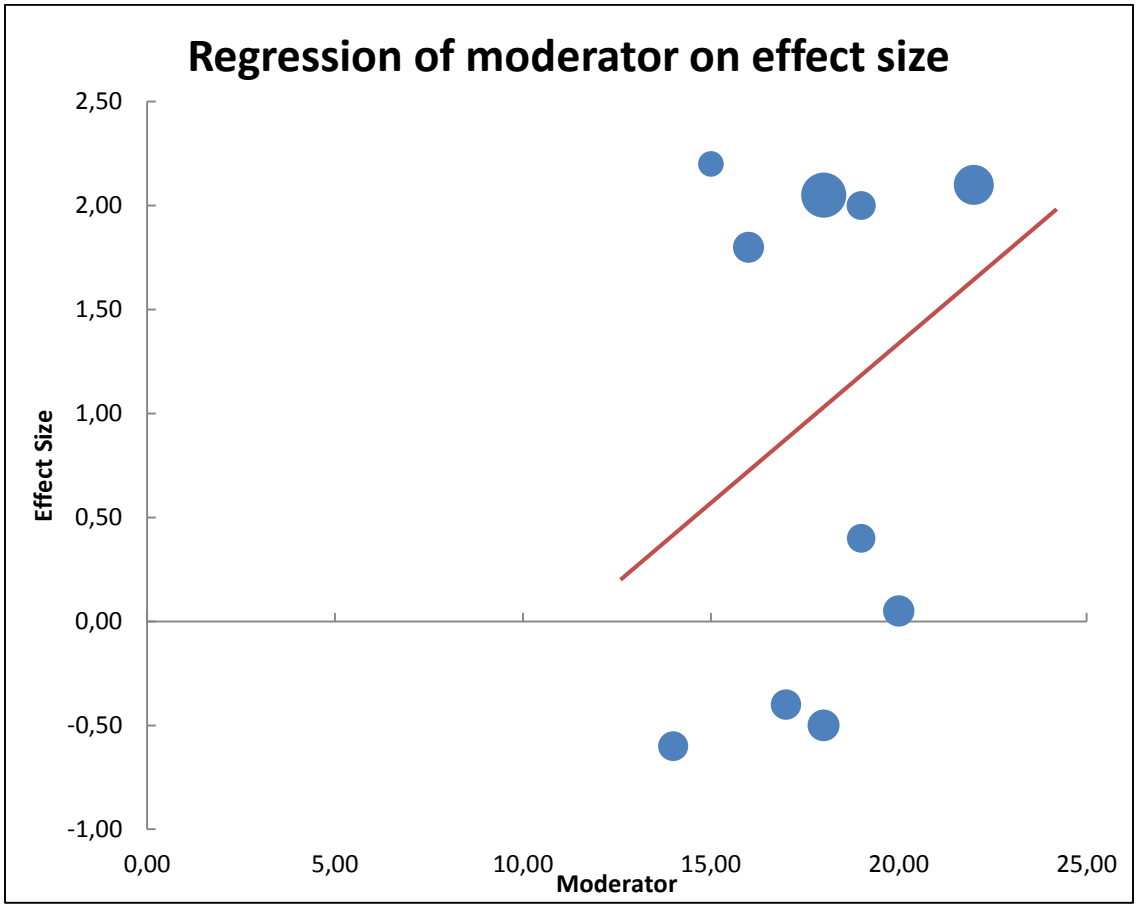


Figure 8: Example of the Moderator Analysis sheet

We know that this is "caused" by the removal of two studies in the top left corner of the plot. However, how sure can one be that the absence of studies in the bottom right corner is not caused by selection bias in the studies that have been conducted? The *p*-value of the regression analysis (not presented in Figure 8) indicates that the regression weight is statistically "not significant". This, however, is to be expected with the (inevitable) low statistical power of this analysis. Lack of significance certainly does not mean absence of a moderating effect! For this reason, information about the statistical significance of this regression result should not be used.

More importantly, because it is the aim of the moderator analysis to explore a potential explanation for dispersion of observed effect sizes, it should be used only in a set of studies in which such dispersion exists. In the preceding subgroup analysis it was observed that subgroup AA (e.g., the United Kingdom) is rather homogeneous. There is no dispersion of effect sizes to be explained in this subgroup (i.e., the group of studies above the regression line in Figures 8 and 9). Therefore, it is not wise to include this subgroup in the moderator analysis. In this example moderator analysis should be performed for subgroup BB (other countries than the United Kingdom) only. Figure 9 presents the results of that analysis.

Figure 9: Example of the moderator analysis in subgroup BB only

The steep regression slope in Figure 9 suggests that higher effects might be observed in other countries in situations in which average class size is higher than in the situations that have been studied. But we cannot be sure whether that is true because no studies have been done in such situations with higher average class size, perhaps because such large average class size (over 25) happens to be non-existent. If this is indeed the case, the conclusion can be drawn that such a potential effect is not relevant for current policy contexts.

The regression weight in subgroup BB (Figure 9) is statistically highly significant (p=0.004) despite the very low statistical power. But, again, the researcher knows this already from the scatter plot. It is clear that the addition of just one study in the top left corner of the scatter plot or, for that matter, in its bottom right corner, could change the results of this analysis considerably. Therefore, the general recommendation for both the subgroup analysis and the moderator analysis is to use its results only for exploration, i.e., for the development of ideas, and not for "testing".

The worry underlying the recommended caution in interpreting the results of a meta-analysis in general, and of the subgroup and moderator analysis in particular, is that the set of studies that are meta-analysed is always an arbitrary selection from all studies that could have been performed about the effect. The selection of this arbitrary set is likely to be biased, but the type and size of this bias is unknown.

5 Publication bias analysis

The set of studies undertaken in a field of study is likely to be biased in many ways. The best known example of such selection bias is the fact that almost all psychological experiments are studies of convenience samples of students. It is very unlikely that experiments with normal people would generate similar effect sizes. One reason for low levels of heterogeneity in psychological experiments might be that they all study the same type of person. This potential selection bias cannot be detected in a meta-analysis of the studies that have been conducted, for the simple reason that a meta-analysis can only draw conclusions from its input and not from what is absent in that input. Because a meta-analytic result is always only a result for the set of populations for which effect size estimates are included in the analysis, the researcher should always be very reluctant in drawing conclusions for a domain.

One form of selection bias that might be detected in the meta-analysis itself is publication bias. In a publication bias analysis it is assumed that the domain is homogeneous, implying that there is no bias even if though the selection of populations for studies is arbitrary. Publication bias analysis is concerned with selection bias that might occur after studies have been conducted, namely when some studies are published and others are not published. It is hypothesized that the chance that a statistically significant result is published is higher than a statistically non-significant result. Hence, the combined effect size in the study might be larger than it is in reality. The publication bias analysis is aimed at (1) signalling this potential publication bias and (2) adjusting the estimate for the combined effect size. Note that this

makes sense only in a homogeneous set of results in which the combined effect size can be interpreted as an estimate of a true effect size in the population.

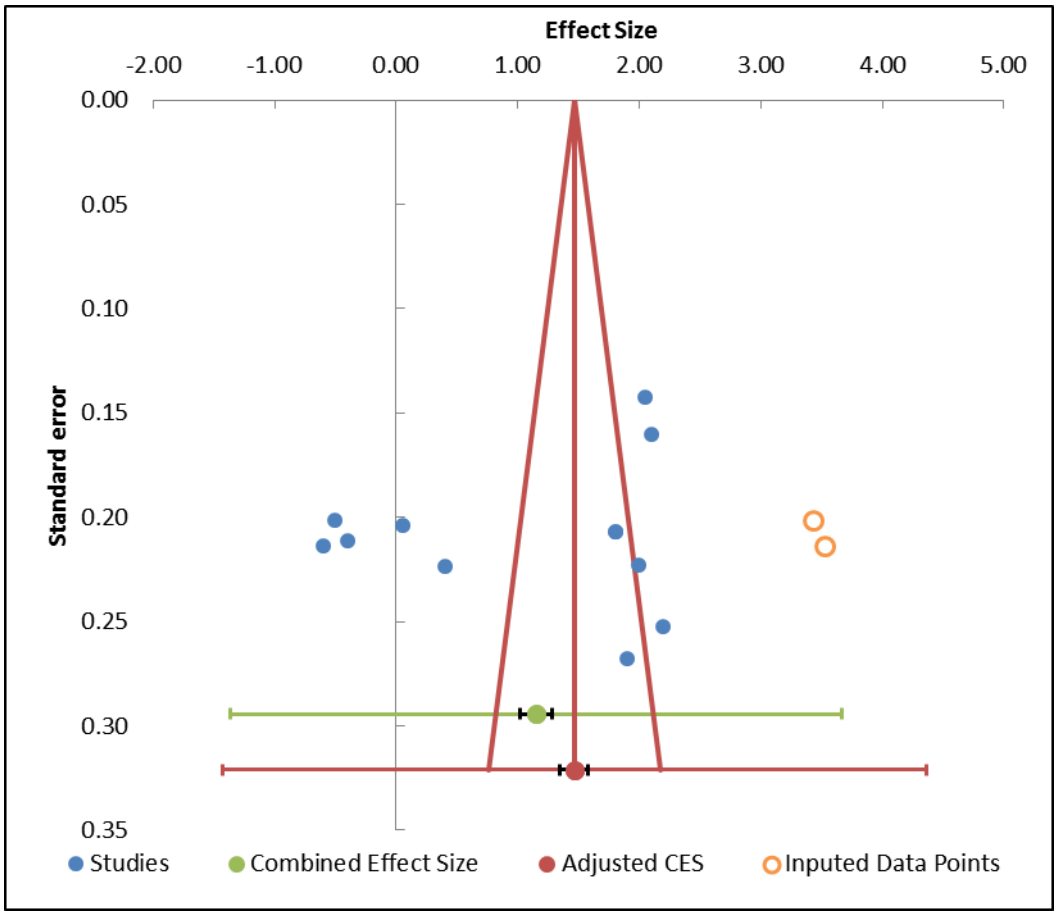


Figure 10: Example of a funnel plot

Meta-Essentials offers six different analyses that might indicate publication bias. One analysis is a funnel plot. It is assumed that observed effect sizes with similar precision (i.e., with similar standard error) should be more or less symmetrically distributed around the combined effect size. As mentioned above, it is hypothesized that there will be more results far from the null than closer to the null. This is not the case in the example (Figure 10). The funnel plot suggests that results with rather large effect sizes are missing. Two results are imputed in order to "adjust" for this absence.

This, again, is an example that illustrates that results of procedures in a meta-analysis, hence also in *Meta-Essentials* should be interpreted with much caution. In this specific example, the result of the funnel plot cannot be interpreted because of the high level of heterogeneity in this set of effect sizes. Publication bias analysis should be performed in a set of homogeneous results only. Hence, in the example that is used in this text so far, publication bias analysis is only useful in subgroup AA.

Figure 11 is the funnel plot for subgroup AA. It suggests that a study with a relatively high effect size is missing. Hence, the adjusted combined effect size (which can be interpreted as an estimate of the true effect size in the subgroup) is a little bit higher than the observed combined effect size. We would expect to find that a study might be missing on the left side of the plot and that adjustment would result in a bit smaller combined effect size. (Remember that the data set is fictitious.)

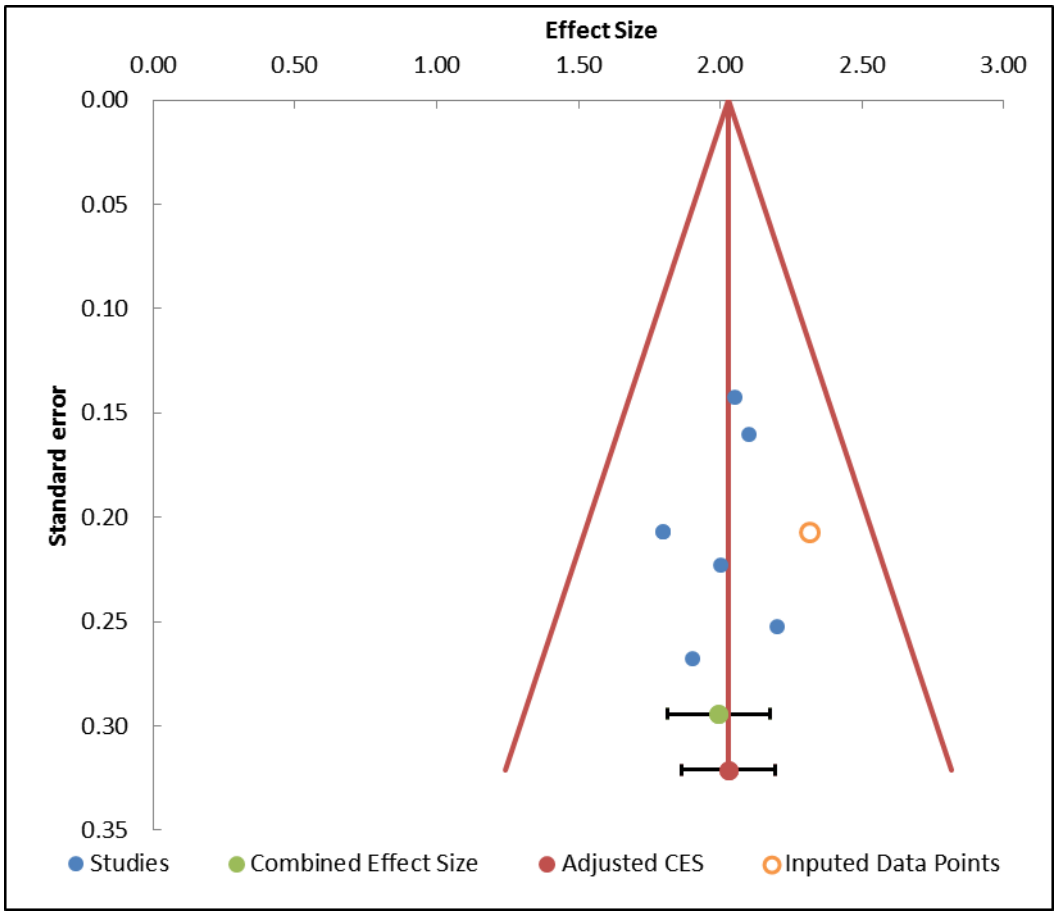


Figure 11: Funnel plot in subgroup AA

The same caution should be applied in the interpretation of the other five estimates of publication bias that are generated in *Meta-Essentials*. None of them has a meaning when generated in a heterogeneous set of studies and, hence, they should only be performed in subgroups of which it is known (through a subgroup analysis) that they are homogeneous. Furthermore, these other five estimates have only diagnostic value – they might help to discover publication bias – but do not provide for an adjusted combined effect size.

6 Conclusion

Meta-analysis is a statistical tool that supports a synthesis and evaluation of the results of studies about an effect that have been conducted and published, and that after publication have been retrieved, read, and critically evaluated by the researcher. All results of the meta-analysis pertain to the set of results that happen to be generated and retrieved. These results are subject to an unknown degree of selection bias which inevitably results from the arbitrary way in which populations (or "samples") are selected for the separate studies. Often, some of the more glaring biases can be known, e.g., when a research literature mainly consists of results of experiments with students and has not covered effects in "real life". Therefore, the first part of any interpretation of meta-analytic results should be an explicit statement about populations that are not yet covered by empirical research and in which different effects might have been observed. Recommendations might be formulated for further research in yet not researched populations.

Meta-analysis generates estimates of a (weighted) average effect size, of the dispersion of effect sizes, of the homogeneity (or heterogeneity) of the total set of observed effect sizes and of subgroups, and supports the exploration of the relevance of potential moderators. Before conclusions are drawn, the degree of heterogeneity should first be assessed and interpreted. "Combined" effect sizes should only be used as an outcome if homogeneity of a group or subgroup of observed effect sizes is without doubt and, even then, only for the domain that is defined by this specific group of populations.

Because relevant heterogeneity is normally found in the social sciences, the main result of most meta-analyses is an insight in the dispersion of true effects. In those cases, meta-analysis functions as a tool for generating hypotheses about "moderators" of the effect.

Meta-analysis should not be used for "testing", and it should not be used for generating statements about the size of an effect in not yet researched parts of the domain or in the entire domain.

In sum, conclusions from a meta-analysis may take the following form:

> Studies of this effect (X → Y) have been conducted in the following populations. The following (types of) populations have not been studied.
> Observed effects range from … to ….
> Effects in subgroup A … range from … to … This means that if X increases with Δx in a population of this subgroup, then Y might increase with at least Δy.
> Effects in subgroup B … range from … to … This means that if X increases with Δx in a population of this subgroup, then Y might increase with at least Δy.

7 Reference

Michael Borenstein et al. (2009), *Introduction to Meta-Analysis*, Chichester (UK): Wiley.