INTRODUCTION TO META-ANALYSIS

by Simon Moss

Introduction

According to some researchers, after PhD students deliberately skim a variety of abstracts rapidly, they can write more fluently. Their productivity improves. To ascertain whether this belief is true, you could

* uncover all the studies that have explored whether rapid reading enhances fluency
* derive the key statistics from these studies, such as the means or averages of each condition
* integrate these statistics to generate one key number—a number that represents the extent to which rapid reading enhances productivity.

In essence, this example typifies a meta-analysis. A meta-analysis is a statistical technique, utilized to integrate all the studies that have tested some intervention or explored some question. The meta-analysis will generate an index called an effect size—a measure that, in essence, indicates the extent to which some intervention or treatment was effective.

|  |  |
| --- | --- |
|  | **Pre-requisites** |

Researchers tend to precede meta-analyses with systematic reviews. Therefore, before you learn about meta-analyses, you should first read, or at least skim, information about

* systematic reviews: available on the CDU website, in the section on how to begin your thesis
* effect sizes: available on the CDU website, in the section on choosing research methodologies

## Example of key excerpts

To learn about meta-analyses, first skim the following table—a table that illustrates some of results of a meta-analysis. The first column presents the headings or topics. The second column presents the details

|  |  |
| --- | --- |
| Section | Example |
| Effect size for the primary outcome | * We generated an estimate of the standard mean difference—the mean difference between the intervention and control conditions, divided by the pooled standard deviation—from 20 studies, using a random effects model * On average, the standard mean difference across these studies was 0.24 (95% confidence interval .13 to .36) * Because this confidence interval does not include 0, these findings indicate the effect of rapid reading on fluency in writing significantly exceeded 0 |
| Effect size for the secondary outcome | * Table 1 presents the standard mean difference for all secondary outcomes, such as the measures of innovation and motivation |
| Funnel plot | * Figure 1 presents a funnel plot. * The funnel plot represents the association between the standard mean difference—a measure of effect size—and standard error of the effect size—a measure of precision—for each study. |
| Assessment of heterogeneity | * The funnel plot shows the effect size varied appreciably across studies * To gauge whether this variability can be ascribed to sampling biases in the study or true variability across studies, we calculated two measures of heterogeneity: Cochrane’s Q and the I2 statistic * For the primary outcome, Cochrane’s Q indicated the true effect sizes differed significantly across the studies, Q(12) = 21.75, p < .05. * Furthermore, the I2 statistic was .39, indicating that 39% of the observed variance in these effect sizes can be ascribed to variability in the true effects rather than sampling bias. * Table 2 presents the Q and I2 statistics for the secondary outcomes as well. |
| Moderator analysis | * We then ascertained whether the mean age of students in each study moderates the association between reading rapidly and fluency * We conducted a mixed effects model, in which the effect sizes were deemed as random effects and the moderators were deemed as fixed effects * As this meta-regression showed, the effect size was positively associated with the mean age of students in each study, B = .154, p < .004. |
| Fail safe N | * To assess whether this significant effect of rapid reading on fluency can be ascribed to the possibility that many non-significant results were overlooked, we applied the Rosenthal fail-safe N test * The Rosenthal fail-safe N was 318, indicating that over 300 studies, with an effect size of 0, would need to be added to convert the significant effect size across studies to a non-significant effect size. * Furthermore, the overall z score was 35.21, p < .001, indicating the effect size would still be significant even if overlooked studies had been included * Table 2 also presents the fail-safe N for the secondary measures as well |
| Begg and Mazumdar Rank Correlation Test | * We also conducted the Begg and Mazumdar Rank Correlation Test to gauge the correlation between effect size and sample size across studies * Typically, a positive correlation indicates that researchers may have been less inclined to publish studies with a small sample that generated a non-significant result, revealing a potential bias in the sample of studies * For the primary measure—productivity—Kendall's tau b = -.143, p = .12, one-tailed, p = .24, two-tailed, indicating no evidence of publication bias * Table 4 presents the results of this test for the secondary outcomes. |
| Egger test | * Furthermore, to assess publication bias, we conducted the Egger test * In particular, we conducted a regression analysis to explore the association between difference between the means over standard error and one over standard error * If the sample of studies is unbiased, the intercept should be 0 * In this instance, the intercept did not differ significantly from 0, B0 = 0.32, (95% confidence interval -. 64 to 1.23), t(18) = 0.42, p > .05 |

## Key complications

The essence of meta-analysis is simple: To conduct a meta-analysis

* for all the studies you collated, estimate the effect size—perhaps using the Cohen’s d formula
* calculate the average of these estimates

Nevertheless, you need to be aware of several complications. This section introduces you to these complications.

**Effect sizes**

For the majority of meta-analyses, the researchers collate studies that compare two conditions, such as a sample of participants who skim many abstracts rapidly and a sample of participants in the control condition. For these meta-analyses, the most common measure of effect size—that is, the measure that gauges the extent to which the treatment is effective—is the standardized mean difference, also known as Cohen’s d. The following table illustrates how to calculate this measure

|  |  |
| --- | --- |
| Step | Example |
| Extract the means and standard deviations of each condition | * For the treatment condition, mean = 7.76, sd = 4 * For the control condition, mean = 6.00, sd = 3 |
| Square the standard deviation of each condition | * 16 and 9 respectively |
| Sum these two answers | * 25 |
| Divide this sum by 2 | * 12.5 |
| Square root this answer—to generate the pooled standard deviation | * 3.54 |
| Divide the difference between the two means by this pooled standard deviation | * 1.76/3.54 = 0.5 |

However, besides this standardized mean difference, software packages, such as Comprehensive Meta-Analysis, apply many alternative formulas to estimate effect size. The following table summarizes some of these alternatives.

|  |  |
| --- | --- |
| Other measures of effect size | Details |
| Bias-corrected standardized mean difference or Hedge’s g | * Similar to the standardized mean difference, but more applicable if the sample size differs considerably between the two groups * Both this measure and the standardized mean difference tend to overestimate the actual effect size—the effect size you would have generated had you utilized the entire population   To compute Hedge’s g, instead of the pooled standard deviation, utilize an adjusted version that is sensitive to the sample size of each condition   * Like the standardized mean difference, square the standard deviation of each condition * Then multiply each of these two numbers by the sample size in this condition – 1 * Add the two answers * Divide the answer by the total number of participants minus 2 |
| Odds ratio | * Applied when you want to assess the extent to which some characteristic affects a categorical outcome, such as whether or not individuals have been diagnosed with a disease |
| Hazard ratio | * Useful for survival analyses, when you want to assess whether the likelihood of death or some other event differs between two conditions |

**Heterogeneity**

Obviously, not all studies in the meta-analysis will generate the same effect size. However,

* some of this variability can be ascribed to sampling bias: That is, if the researchers had recruited every participant in the population, these studies might have generated the same effect size
* alternatively, even if every researchers had recruited every participant in the population, these studies might have generated diverging estimates of effect size; That is, the actual or true effect size might have varied across studies—called heterogeneity.

To gauge heterogeneity, researchers tend to calculate two indices: the Cochrane Q test and I2 statistic. To illustrate, if the true effect size is the same across studies, the Cochrane Q value should approximate the degrees of freedom—or number of studies minus 1. Furthermore, the I2 statistic is the percentage of variance in the effect sizes that can be ascribed to variability in true effects rather than sampling error.

**Moderation**

If the effect sizes are heterogenous and thus vary markedly across studies, researchers may want to assess whether some of other characteristic affects these effect sizes. For example

* perhaps rapid reading is more likely to enhance fluency in America than in Europe
* hence, location could affect or moderate the effect size
* in this example, location is labelled a moderator

The most common technique to examine moderators is called meta-regression. In essence, the researcher conducts a linear regression, discussed in another document on the CDU website, to ascertain whether the moderator is associated with the effect size. This technique, however, is effective only if the number of studies is adequate, such as 50 or greater.

**Fail-safe N**

One problem complicates a sizeable proportion of meta-analyses: Often, researchers do not publish non-significant results. Consequently, researchers who conduct meta-analyses may inadvertently overlook studies that generated low effect sizes. The studies in their sample, therefore, might be biased. That is, these studies might overestimate the true or average effect size.

To assess this problem, many researchers calculate the Fail-safe N. In particular,

* The Fail-safe N represents the approximate number of non-significant or overlooked studies that could have changed the results.
* For example, if the Fail-safe N is 5, a low number, five overlooked studies could have shifted the results from a significant effect to a non-significant effect
* Consequently, the researchers cannot be confident the effect—such as the impact of rapid reading on fluency—is really significant: A few overlooked studies could have reversed the results
* Several methods have been developed to compute the Fail-safe N, including the classic or Rosenthal (1979) method, the Gleser and Olkin (1996) method, and the Orwin (1983) method.

**Measures of publication bias**

Rather than only compute a fail-safe N, some researchers also explore other patterns in the data, especially the funnel plot, that might imply the sample of studies are biased and not representative of all the studies on this topic. The following table clarifies the rationale of these procedures.

|  |  |
| --- | --- |
| Technique | Rationale |
| Funnel plot | * Researchers might be reluctant to publish studies in which the results were not significant and the sample size was small * In contrast, researchers may still publish studies in which the results were not significant if the sample size was large; the reason is that such results cannot be as readily ascribed to a small sample and thus may not be as dismissed as readily * If this premise is true, a small sample size should coincide with larger effect sizes * The funnel plot represents the correlation between the effect size and sample size of each study * If the studies are unbiased, effect size should be unrelated to sample size * Consequently, this plot should be symmetrical     **Variants**   * Some researchers represent the correlation between the effect size and standard error of each study—a very similar plot because standard error is primarily a function of sample size * Other variants are sometimes reported; each is similar but some plots underscore specific deviations from symmetry more effectively |
| Begg and Mazumdar Rank Correlation Test | * As indicated above, if the sample of studies are biased and not representative of all studies on a topic, a small sample size should coincide with larger effect sizes * The Begg and Mazumdar Rank Correlation Test measures the association between sample size and effect size * A significant negative correlation indicates that researchers were indeed reluctant to publish studies in which the results were not significant especially if the sample size was small * This negative correlation thus implies that non-significant results had been overlooked. * One drawback of this technique, however, is the test is not powerful when the number of studies is lower than 25 or so. |
| Egger test | * Furthermore, to assess publication bias, researchers often conduct the Egger test * In particular, this test utilizes regression analysis to explore the association between difference between the means / standard error and 1/ standard error * If the sample of studies is unbiased, the intercept should be 0 |

**Corrections for publication bias: Trim and Fill**

If researchers do uncover a publication bias, they might attempt to correct these biases. That is, they attempt to estimate what the effect sizes would have been had their sample of studies not been biased. One of the most common approaches is called Duval and Tweedie's Trim and Fill test. In principle, the software

* eradicates a subset of studies to generate a symmetrical funnel plot
* estimates the effect size from this subset of studies—called the estimated effect
* assesses the level of asymmetry to estimate the number of overlooked or missing studies
* replaces the eradicated and overlooked studies with numbers that equal the estimated effect on average

**Sensitivity analyses**

Many researchers also conduct sensitivity analyses. In essence, the researcher, or at least the software, repeats the meta-analysis as many times as the number of studies in the sample. Each time, the meta-analysis computes the overall effect size without one of these studies. These analyses determine the extent to which each study influences the estimated effect size.

**Fixed versus random effects**

Researchers need to decide which of two models they would like to utilize: fixed effects or random effects. The following table clarifies the difference between these alternatives.

|  |  |  |
| --- | --- | --- |
|  | Fixed effects | Random effects |
| Assumption | Assumes the true effect is the same in each study. If researchers had utilized everyone in the population, precluding sampling error, every study would generate the same effect size | Does not assume the true effect is the same in each study |
| Applicability | Appropriate if the population utilized in each study was the same. For example, if all the studies were restricted to research students in Australia, a fixed effect might be applicable | Appropriate if the population varied across the studies—the more prevalent circumstance |
| Implications | If a random effect model is suitable, but the researcher assumes a fixed effect, the confidence interval and thus precision is underestimated | If a fixed effect model is utilized, but the researchers assumes a random effect, the consequences are trivial; a test of heterogeneity will show the true effect does or does not vary across studies |

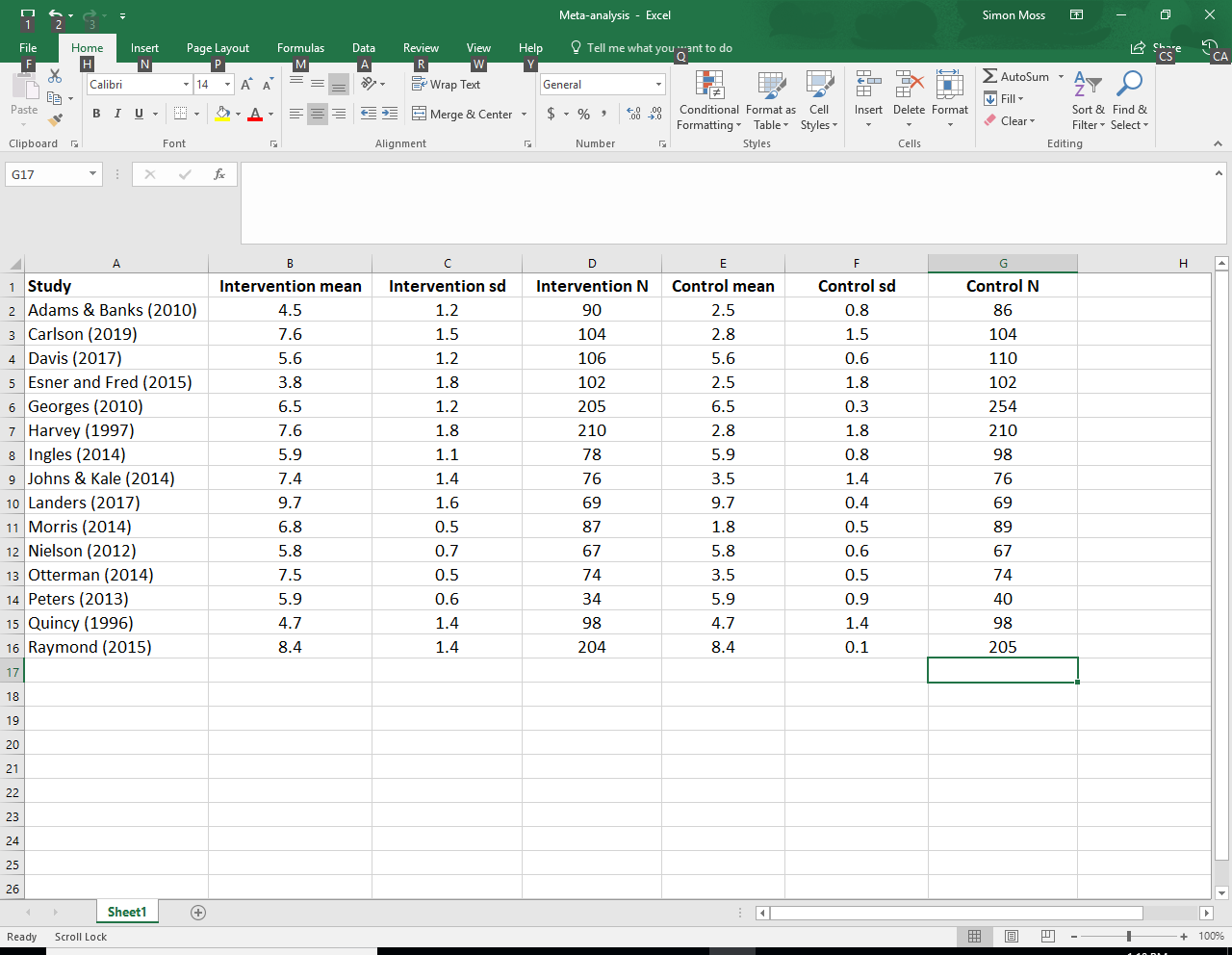
## How to conduct the meta-analysis

In practice, once you have collated all your studies, meta-analysis is not too hard to undertake. Specifically, most researchers utilize software, such as a program called Comprehensive Meta-Analysis, to calculate the statistics. This section offers some guidance on how to utilize this software.

**Step 1**. **Extract the data into Excel**

Construct a spreadsheet, perhaps in Excel or using similar software, that resembles the following example. That is

* each row represents a separate study
* the first column labels the study—typically with the authors
* the next set of columns specify the mean, standard deviation, and sample size of the intervention and control groups separately.



Some complications might occasionally unfold. For example

* some studies might not report the mean, standard deviation, and sample size of each condition
* in these instances, construct columns to represent the other reported statistics, such as p values
* some studies do not merely report two conditions.
* To illustrate, if you want to include more than two interventions in your meta-analysis, you need to learn about a variant called network meta-analysis—a slightly more complex technique

**Step 2.** **Download the latest version of Comprehensive Meta-Analysis**

To download this software

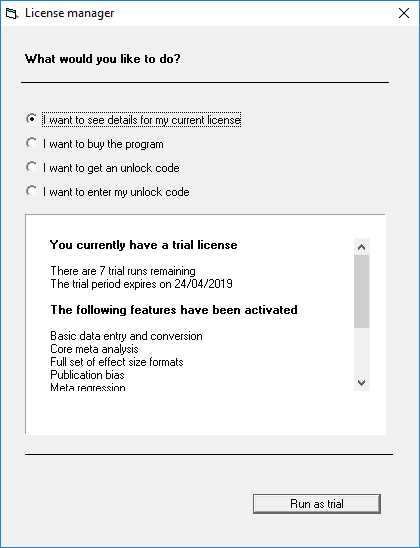
* Visit <https://www.meta-analysis.com/>
* Press the “Download” tab to uncover various options
* You could choose the CMA free 10-day trial. Or you could choose to purchase the software
* The following table presents the benefits of the free and paid versions
* Click the option you prefer and follow the instructions to download
* The software does not work on Macs, unless you have installed a Windows emulator

|  |  |
| --- | --- |
| Drawbacks of the free version | Drawbacks of the paid version |
| Lasts only 10 days—not a major concern if you have already entered the data into a spreadsheet | Costs about $250 |
| Excludes some tests and statistics that are available in the paid version. However, you might be able to utilize other software or formulas to calculate these statistics |  |
| The funnel plot appears with a label that indicates you are using only a free version. For your thesis, you would need to construct the Funnel plot yourself, perhaps in Excel |  |

Perhaps you could practice with the free version. But, when you are ready to proceed, you should purchase the paid version.

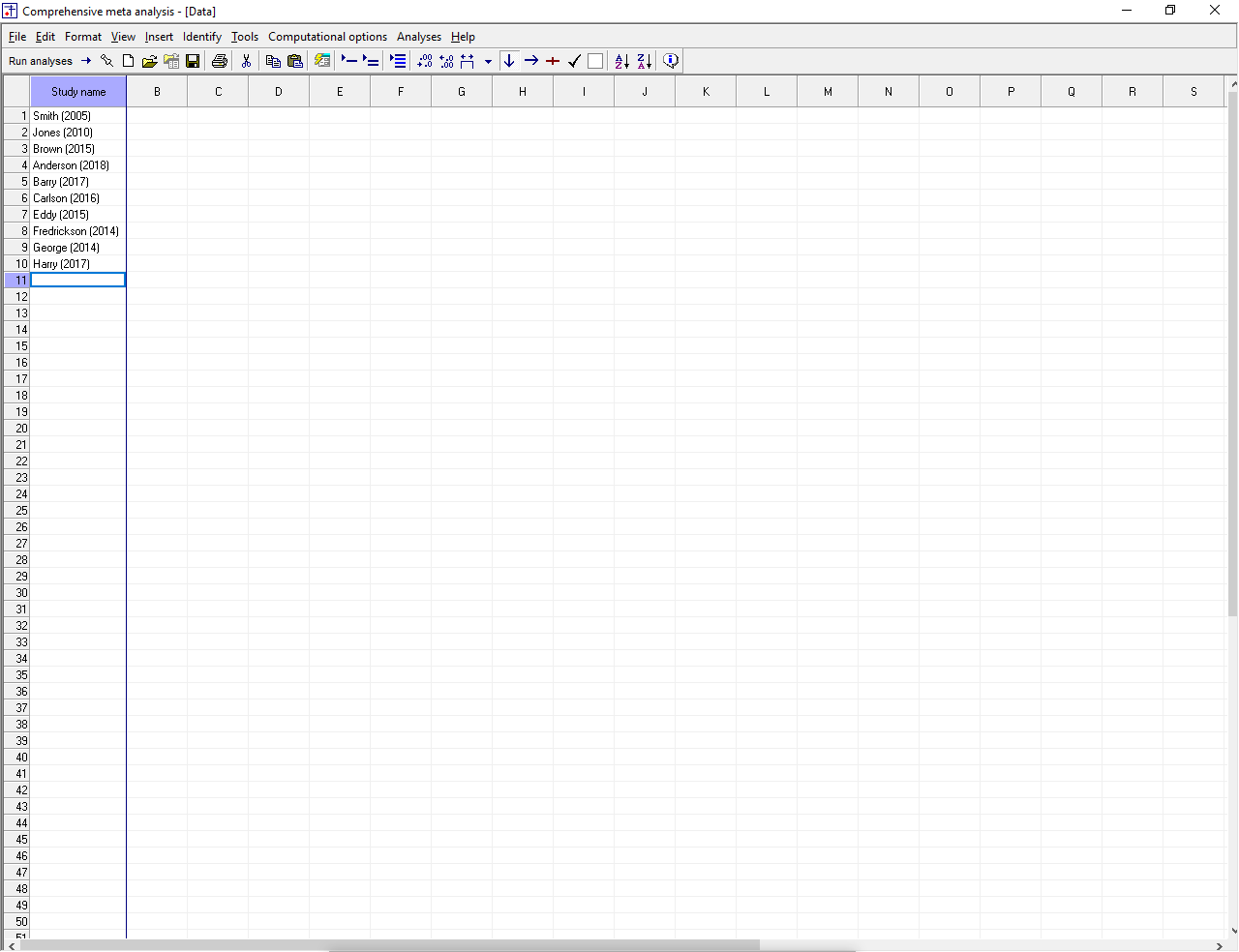
**Step 3. Enter the study labels**

After you open the software, you might receive the following screen. Simply Press “Run as trial” on the bottom. Then, press “Close” if you receive a Welcome message



Your first step is to enter the study labels into the first column. To achieve this goal

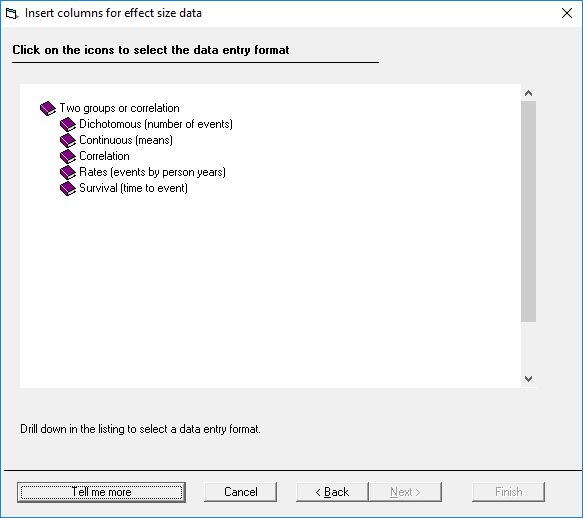
* Click the “Insert” menu and then “Column for” and finally “Study names”
* A new column should appear in the spreadsheet called “Study names”
* Merely copy and paste the study names from Excel into this column. Some computers experience a bug in which the paste does not work.
* This procedure should generate this screen



**Step 4. Enter the study statistics**

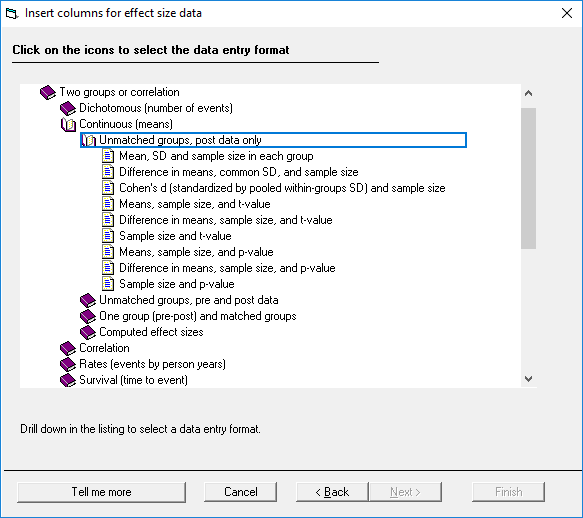
You now want to enter the means, standard deviations, sample sizes, and other measures of each condition into the spreadsheet. To achieve this goal

* Click the “Insert” menu and then “Column for” and finally “effect size data”
* In general, choose “Next” and “Next”, because the default options are often the most applicable, again to generate the following screen



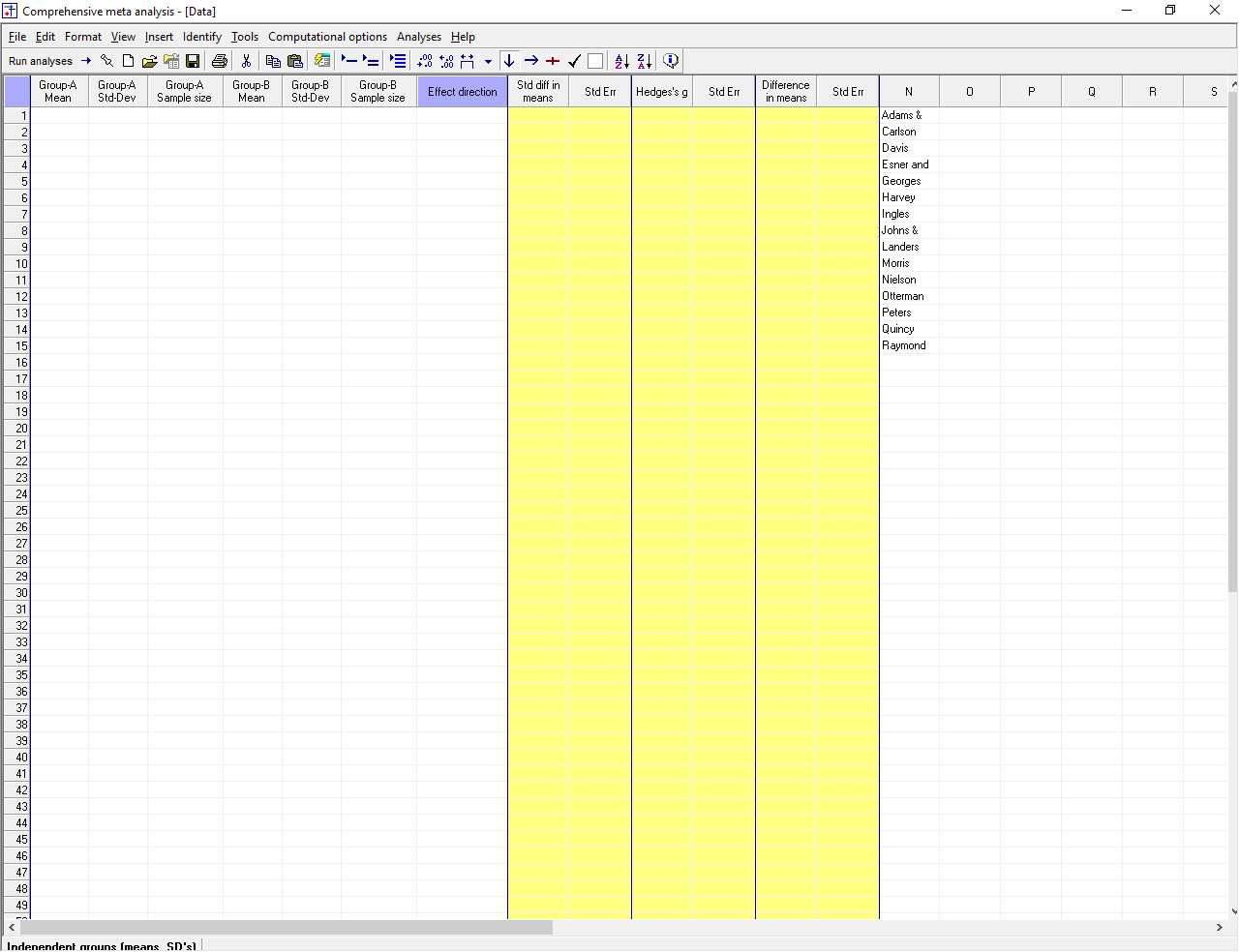
Now, if you want to compare the groups on some numerical measure, such as number of words written during one day

* choose “Continuous”, generating a set of other options
* then choose “Unmatched group, post data only”, to generate the following screen



You now need to inform the software of which data you have extracted. In this instance, assume you have extracted the mean, standard deviation, and sample size of each condition—a common situation. Accordingly

* Double click this top option
* The software will generate the relevant columns



You can now

* Label Group A and Group B, such as “intervention” and “control”; the computer will prompt you to specify these labels
* Copy and paste the Excel data into the first six columns in this spreadsheet
* SPSS will then calculate information in the yellow columns, such as the Standard Difference in Means—a measure of effect size.

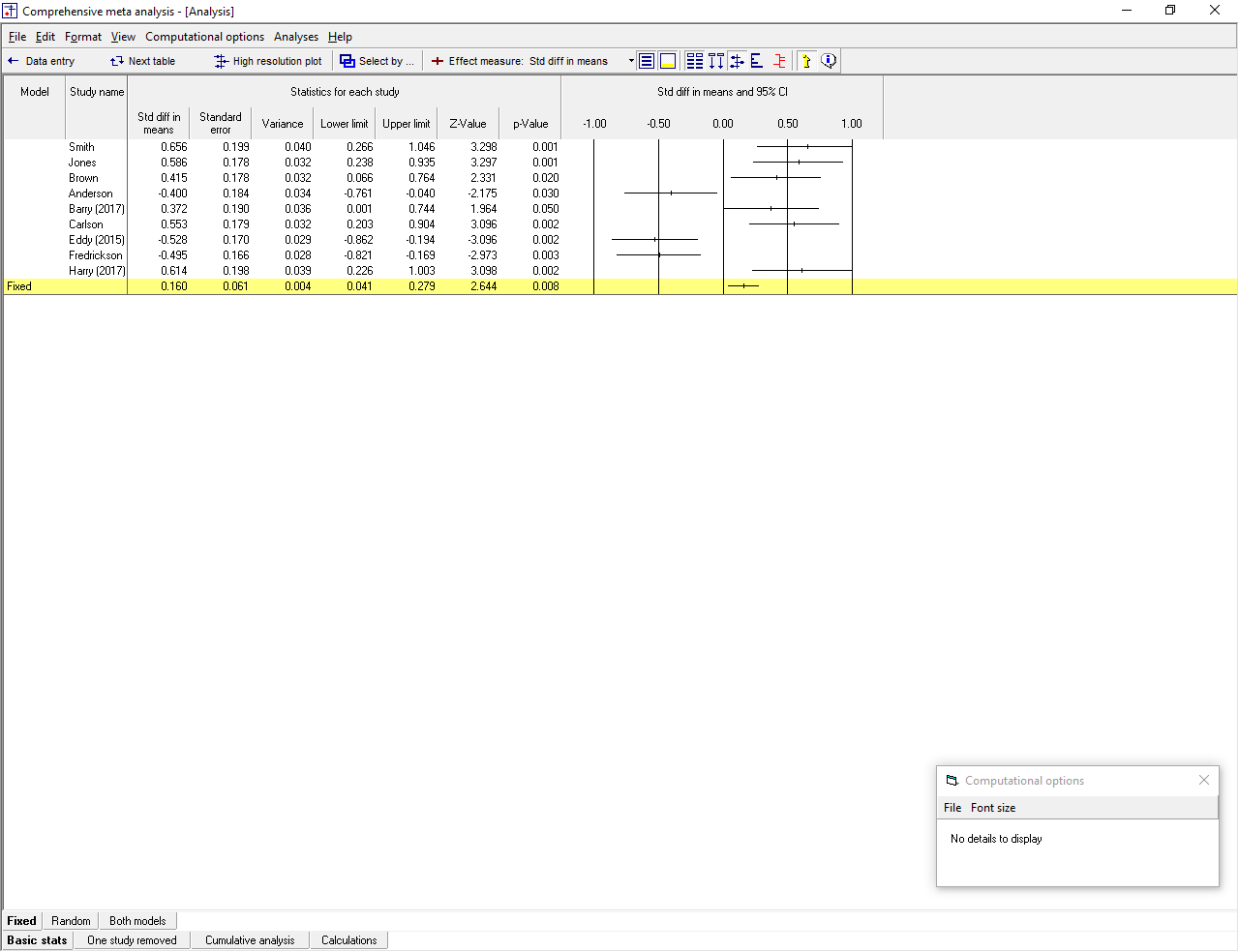
**Step 4a. Refine the data entry**

The previous example illustrated a common, but not universal, circumstance. The following table specifies some other complications you might need to consider

|  |  |
| --- | --- |
| Complication | Details |
| * Your design is different—and you do not merely want to compare two conditions on some numerical variable | * While completing the previous steps, you might need to choose alternative book icon |
| * Some of the studies present other statistics, such as a p value and standard error instead of the mean, standard deviation, and sample size of each condition | * While completing the previous steps, you might need to choose alternative article icon |
| * For some studies, the outcome measure was in the reverse direction. For example, for one study, low scores could represent positive outcomes | * In the column called effect direction in the spreadsheet, usually choose auto * However, sometimes auto will generate unsuitable measures of effect size—such as standardized mean difference * For example, perhaps a positive standardized mean difference usually implies the treatment effect outperformed the control condition * But, if the outcome measure was reverse scores, a negative standardized mean difference might imply the treatment effect outperformed the control condition * Therefore, you might have to shift the sign of these standardized mean difference as well as the other effect sizes—to ensure they reflect whether the intervention or control generated favorable outcomes * To achieve this goal, specify positive or negative in the column called effect direction |

**Step 5. Conduct the analyses**

To calculate the various statistics and plots, simply press “Run analyses”—an option on the far left. This procedure will generate a screen the resembles the following example.



To illustrate

* The final row, called Fixed, presents an aggregate of all the studies.
* In this instance, the standardized difference in means across the studies is 0.16—usually regarded as a small effect size
* But the p value is .008, indicating the effect size significantly differs from 0
* If you press some of the additional buttons, such as “Next table” or “High resolution plot”, more information, such as tests of heterogeneity appear.
* To apply a Random effects model instead, you can choose a button on the bottom left—not shown here.

**Step 6. Learn more about this program**

This discussion thus far was designed to familiarize you with the software. Once familiar with the software, you can readily learn about the various features and complications. To learn about the software,

* Visit <https://www.meta-analysis.com/pages/videotutorials.php>
* Usually about nine videos will appear, under the headings data entry, basic analyses, and advanced analyses.
* Watch the first video under each heading

**APPENDIX 1: MEGA-ANALYSIS**

Mega-analysis is a variant or alternative to meta-analysis, suitable whenever researchers can access the original data of all studies. To learn about this technique, note that meta-analysis, in essence, comprises two key phases:

* first, the researcher computes the effect size of each relevant study in the analysis
* second, the researcher derives statistics from these effect sizes—such as the average

In contrast, to conduct mega-analysis, researchers do not need to compute the effects size of each study in the analysis (see Burke et al., 2017; DeRubeis et al., 1999). Instead, the researcher

* pools or collates all the data from all the relevant studies into one data file
* derive statistics from this pooled data. The following diagram represents this difference between meta-analysis and mega-analysis.

A picture containing timeline

Description automatically generated

Researchers can benefit from several advantages of mega-analysis over meta-analysis. For example,

* in mega-analysis, you can assess the effects of individual characteristics more effectively
* mega-analysis circumvents some of the assumptions that must be fulfilled to conduct meta-analysis—such as the assumption that effect sizes are normally distributed across studies

The complication is that you need to access the original data of each study, usually by contacting the authors. However, once you have pooled the data, you can apply classical statistical methods, such as multi-level modeling, to analyze the data. Typically, researchers do include a variable that differences the studies. But otherwise, standard techniques are suitable.

**References**

Burke, D. L., Ensor, J., & Riley, R. D. (2017). Meta-analysis using individual participant data: One-stage and two-stage approaches, and why they may differ. Statistics in Medicine, 36(5), 855–875.

DeRubeis, R. J., Gelfand, L. A., Tang, T. Z., & Simons, A. D. (1999). Medications versus cognitive behavior therapy for severely depressed outpatients: Mega-analysis of four randomized comparisons. The American Journal of Psychiatry, 156(7), 1007–1013.