Some Basic Points Concerning Meta-Analysis

DARYL S. PAULSON

ABSTRACT

Multiple studies have been performed on a variety of substances, often producing contradictory results. Meta-analysis has provided a means of evaluating these disparate results, combining them into a summary statistic. Using continuous data for baseline and one sample point, several studies were evaluated to achieve a single result, demonstrating the meta-analysis evaluation process.

ABBREVIATIONS: ANOVA-analysis of variance, H_0-null hypothesis, H_A-alternative hypothesis, FDA-Food and Drug Administration.

INDEX TERMS: Meta-analysis, Statistics, Study results

Clin Lab Sci 2013;26(1):30

***Daryl S. Paulson, PhD**, BioScience Laboratories, Inc., Bozeman, MT*

***Address for Correspondence: Daryl S. Paulson, PhD**, President and Chief Executive Officer, BioScience Laboratories, Inc., 300 North Willson Avenue, Suite 1, Bozeman, MT 59715, (406) 587-5735, ext. 104, dpaulson@biosciencelabs.com*

In the past, it has been very useful to perform statistical analyses on study data to evaluate the test substance's effectiveness. If done correctly, the analysis can determine if the test substance was effective. Over the years, multiple studies have been performed, evaluating the same substances in the same general ways. What does one do when one study says a substance is effective, but another study says that substance is not effective?

Meta-analysis is a statistical methodology that allows one to evaluate studies conducted at different laboratory test sites, at different time periods, by different scientists, on different test subjects, and combine those results into one study.[1,2] Meta-analysis uses the results gained from a number of different studies as its data points and analyzes them.[11]

Meta-Analysis for Continuous Data

Meta-analysis can evaluate continuous data, binary data, or the correlation among data. The focus in this paper is continuous data, rather than binary or correlational data, as it is used more frequently in scientific fields. Let us take hand disinfectants as an example. Researchers evaluated a product by measuring the baseline sample (pre-product application) and a post-product application sample. A baseline value and a post-application sample divided by standard deviation was used to calculate a D value in this work. It equals:

$$D = \frac{\bar{x}_{BL} - \bar{x}_{sampletime}}{s_{pooled}}$$

where:

D = the dependent variable, which is the difference of the baseline minus the sample time divided by the pooled standard deviation. Each of the D values was the result of one complete study. The baselines were different for each study, so subtracting the post-application sample time from the baseline provided the reduction in microorganisms. This procedure adjusted all the studies, making it possible to compare them directly by their reduction values. (We will discuss dividing the reduction by the pooled standard deviation in the s_{pooled} section.)

$\bar{x}_{sample\ time}$ = the $\log_{10}$ count average of the sample time. The population counts were not linear but exponential. This greatly complicated the statistical model; hence, they were transformed into linear scale, by taking the $\log_{10}$ of the plate count data.

$\bar{x}_{BL}$ = the $\log_{10}$ colony count average of the baseline. The same transformation to a $\log_{10}$ scale was applied to baseline data.

S_{pooled} = this study involved the baseline and the post-application sample time microbial counts on the same subject. The hands were selected for baseline and post-application sample according to the randomization schedule (left hand versus right hand). This was a paired test (the same subject was used for both readings), which made the standard deviation a pooled standard deviation. However, each study had a different standard deviation. So, to adjust the data for easy comparison, the reductions were divided by the standard deviation. The end result was the D value, which informs the reader how many standard deviations the reduction (baseline – post-application sample) is. For example, if the average baseline was 5.00 $\log_{10}$ and the average wash was 3.00 $\log_{10}$, then 5.00 – 3.00 = 2.00 $\log_{10}$. That is, the product reduced the microbial colony counts by 2.00 $\log_{10}$. If the standard deviation was 1.00 $\log_{10}$, the 2.00 = 2.00/1 represented the number of standard deviations to the right the reduction represented. The S_{pooled} formula was:

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where:

n_1 = sample size of the baseline data

s_1^2 = variance of the baseline data

n_2 = sample size of the application data

s_2^2 = variance of the application data

Hedge's g Statistic

The D statistic $\left(D = \frac{\bar{x}_{BL} - \bar{x}_{sampletime}}{s_{pooled}} \right)$ was slightly biased in that it over-estimated the differences with small sample sizes. This bias was removed through a simple correction – the Hedges' g calculation.[12]

The Hedges' g statistic was calculated:

$$g = J \times D$$

where:

$J = 1 - \frac{3}{4df - 1}$ (df = degrees of freedom to estimate S_{pooled}). This was the correction factor. The degrees of freedom was the denominator of the standard deviation; $df = (n_1 + n_2 - 2)$.

D = the dependent variable from the outcome of the difference divided by the pooled standard deviation). It represented how large the reduction was in terms of the standard deviation.

The Hedges' g statistic was used in all the calculations.

Fixed or Random Effects

There are two popular statistical models for meta-analysis, fixed and random effects. Recall that in a general statistical model (e.g., Analysis of Variance [ANOVA]), a fixed effects model means that the dependent variables were chosen before a study began. Using the hand disinfectant example – a healthcare personnel handwash with chlorhexidine gluconate as the main ingredient – if a researcher wanted to evaluate the product against the best-selling chlorhexidine gluconate, s/he decided against what product to test. This was a fixed effects model, because the variable (product) was chosen deliberately. On the other hand, if a random product selection was made from all the chlorhexidine gluconate products available, it was randomly selected for comparison.[15]

In meta-analysis, the fixed and random effects mean have completely different meanings.[9] For the fixed effects model, it was assumed that there was one true effect for all studies in the model. In other words, a drug's effect had the same value among all the different studies, and any differences were purely sampling error.

For the random effects model, it was assumed that there was not one true average value for all the studies combined. They were different values.

In summary, the fixed effects model handled the data with these assumptions:

- There was one true effect size for all studies
- All the different effects were actually sampling errors
- Weights were assigned high values for studies

that had high sample sizes,
- Small sample size studies were assigned smaller weights, and
- Weights of study = $\frac{1}{V_y}$, where $V_y = s^2$ = variance

For random effects, they were handled differently:
- There were different true effects sizes, depending upon the study
- The different effects sizes were presumed real, not sampling errors
- Higher weights were not assigned to studies with larger sample sizes,
- Small sample size studies did not get smaller weights, and
- Weights of study = $\frac{1}{V^*_y}$, where $V^*_y = s^2 + \tau^2$, s^2 = variance, τ^2 = within-subject variance

Understanding these effects is very important in meta-analysis, for using them changes the confidence intervals, as well as the grand total value, often dramatically.

The statistics are rather tedious to compute and usually performed by a computer software program, so they will not be discussed further. For a background of how different statistical programs are run, the procedures can be reviewed.[1,3]

Which Model Should Be Used?

There are several thoughts about which model – fixed or random – should be used. The first paradigm is that the main effects of this hand disinfectant example should be random effects, because the studies included were performed at different times by different people using different subjects. The results were expected to be different. The second paradigm states that it makes sense to use fixed effects if two conditions are met: first, if the researcher believed that all the studies were identical; and second, if the goal was to compare a common effects size from identical populations.[3] These two conditions are not common in hand disinfectant studies. First, for these types of studies where media was placed on the hands, the initial population probably varied, providing different baselines among the different studies. However, this was corrected by using the reduction (baseline – post-application sample) value. This part of the *D* value was discussed previously. For studies that used populations of bacteria normally living on the hand surfaces, and the subjects' counts dependence upon time of year, humidity, and temperature, this method may also be used. Think about the many other areas that are studied, and you will probably see similarities.

The second question was "will the sample size be consistent among studies?" Some studies had as few as five subjects, and others had more than 100; so they varied. To be safe, use of the random effects model was suggested.[9]

These two paradigms were not discussed completely. To determine if the fixed or random effects should be used, there are several other valuable tools. For example, a researcher can check if the groups appear homogenous (the same) or heterogeneous (different). This test examines the *Q* values (discussed later). There are also other tests like finding the I^2, values, which is a kind of signal to noise ratio test, and the *T* value is another test, which is the standard deviation of the true effects size.

These were not the only factors of concern with this study, as was discovered later when the subgroups – the application times (30 seconds and 1 minute) – were included in this handwash model. These applications times were consistent no matter what product was tested or by whom. This categorized the subgroup as "fixed effect," which shall be discussed later. However, had the researcher discovered differences in the studies being compared – for example, if the application times varied inconsistently, the model would have been a random effects component.

Importance of Selecting All Studies, Not Just the "Good" Ones

This is a central point in meta-analysis. It is critical to select all the studies that one can find for the evaluation.[3] Otherwise, for example, the results may be skewed in a direction desirable to the researcher. Using the hand disinfectant example, if the researcher selected only those studies that showed the product to be effective and dismissed those that showed it was not, the meta-analysis would have been biased. But how would a reader know this?

When beginning a meta-analysis, the researcher must define a reasonable inclusion/ exclusion criteria list for the studies and publish it with the results. For example, the inclusion criteria identified all studies that used the FDA handwash guidelines[4] for hand disinfectant studies. Notice that these items were not the way "this test was supposed to be run," but the way the analysis was designed. The exclusion criteria consisted of studies with data generation not clearly understood, types of studies using guidelines different from those of the FDA, and studies not performed in a randomized manner. These two areas require much time and consideration for the selection of studies to be used.[5] Table 1 contains the series of studies included in this analysis.

The eight studies in this meta-analysis fit the inclusion/exclusion criteria just presented.

Meta-Analysis

The main statistical test was:

$$H_0: BL^* = W$$
$$HA: BL \neq W$$

* *BL* = Baseline; *W* = Wash

That is, does a significant difference exist between the baseline and the wash (post-application sample)?

The preliminary meta-analysis is displayed in Figure 1, where each of these studies went through an analysis and received a final or grand total score (bottom line). The 95% confidence intervals are also given, with a probability value. The probability value or *p*-value is the probability (that the true Hedges' *g* value was equal to or greater than $x^* \mid H_0$ true) $\leq \alpha$ or level of significance. x^* = Hedges' *g* actually calculated, and the level of significance for this test is $\alpha = 0.05$.

Each of these eight studies achieved a significance of $p < 0.000$. Note the diamond at the bottom of the graph represents the 95% confidence interval of all eight of the tests (3.920 – 8.285). The grand total of the Hedges' $g = 6.102$. The values were synthesized into one value for the entire meta-analysis. The alternative hypothesis was accepted (H_A); the test was significant. The baseline and the post-application samples were different at $\alpha = 0.05$.

Looking at the graph portion (right-hand side) of Figure 1, differences appeared among the groups, even though this was a random effects study, in which some variation among studies was expected. That is, the Hedge's *g* values did not seem to be homogenous (roughly the same), but instead were heterogeneous (different). So a homogeneity versus heterogeneity test was performed. The random effects model was temporarily changed to a fixed effects model. The test hypotheses were:

H_0: All groups are homogenous, or the same.
H_A: The groups are different (heterogeneous).

Table 1. Study Data

		Baseline			Application of Product		
	Study Name	Group A Mean	Group A Std Dev	Group A Sample Size	Group B Mean	Group B Std Dev	Group B Sample Size
1	12	6.800	0.340	30	3.560	0.265	30
2	13	7.400	0.389	50	3.890	0.367	50
3	24	6.880	0.678	38	4.789	0.567	38
4	45	7.900	0.564	20	4.890	0.452	20
5	67	5.890	0.780	10	3.870	0.959	10
6	71	8.900	0.561	62	6.900	0.780	62
7	26	7.520	0.294	75	3.190	0.379	75
8	48	6.300	0.593	16	4.870	0.362	16

META ANALYSIS

Statistics for each study

Study Name	Subgroup within study	Hedges' g	Standard Error	Variance	Lower limit	Upper limit	z-value	p-value
12.000	1 min	10.491	0.982	0.982	8.549	12.434	10.586	0.000
13.000	1 min	9.211	0.464	0.464	7.876	10.545	13.528	0.000
24.000	30 sec	3.312	0.124	0.124	2.622	4.001	9.415	0.000
45.000	30 sec	5.773	0.513	0.513	4.369	7.176	8.063	0.000
67.000	30 sec	2.213	0.306	0.306	1.129	3.297	4.002	0.000
71.000	30 sec	2.926	0.066	0.066	2.421	3.431	11.356	0.000
26.000	1 min	12.702	0.564	0.564	11.229	14.174	16.910	0.000
48.000	30 sec	2.837	0.245	0.245	1.868	3.807	5.737	0.000
Grand Total		1.114	1.240	1.240	3.920	8.285	5.480	0.000

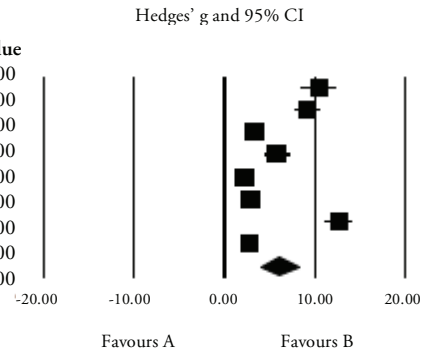

Figure 1. Meta-Analysis

The *Q* value is a measure of the weighted squared deviations. If the *Q* value was quite large (*Q* > *degrees of freedom* {*df*}), the study had greater deviation (*Q*) than was expected. For this analysis, the heterogeneity test (*Q* value) was 278.472, with $k - 1$ or $8 - 1 = 7$ degrees of freedom; *k* is the number of studies evaluated.

Q value	–	degrees of freedom	=	final value
278.472	–	7	=	271.472

Using the Chi Square test with $k - 1$ degrees of freedom and checking 271.472, a significance of less than 0.05 was achieved ($p < 0.000$). The studies were not homogenous. This *Q* value was too large to ignore.

Viewing the data in Figure 1, study names, the study numbers 12, 13, and 26 were different from the other studies in that they appeared to be much more effective.

To get a clearer picture of this and determine the cause, the data were rearranged from high to low and reviewed (Figure 2).

Going back to the original studies to determine if the products were different or if the application times were longer, a difference was noted. It was discovered there were two product application times, 1 minute and 30 seconds, that were not noted at first. A subgroup (time of wash, either 1 minute or 30 seconds) was then included in the model. Then the temporary fixed model was changed back to a random effects model. If these

META ANALYSIS

Statistics for each study

Study Name	Subgroup within study	Hedges' g	Standard Error	Variance	Lower limit	Upper limit	z-value	p-value
26.000	1 min	12.702	0.751	0.564	11.229	14.174	16.910	0.000
12.000	1 min	10.491	0.991	0.982	8.549	12.434	10.586	0.000
13.000	1 min	9.211	0.681	0.464	7.876	10.545	13.528	0.000
45.000	30 sec	5.773	0.716	0.513	4.369	7.176	8.063	0.000
24.000	30 sec	3.312	0.352	0.124	2.622	4.001	9.415	0.000
48.000	30 sec	2.837	0.495	0.245	1.868	3.807	5.737	0.000
71.000	30 sec	2.926	0.258	0.066	2.421	3.431	11.356	0.000
67.000	30 sec	2.213	0.553	0.306	1.129	3.297	4.002	0.000
Grand Total		6.102	1.114	1.240	3.920	8.285	5.480	0.000

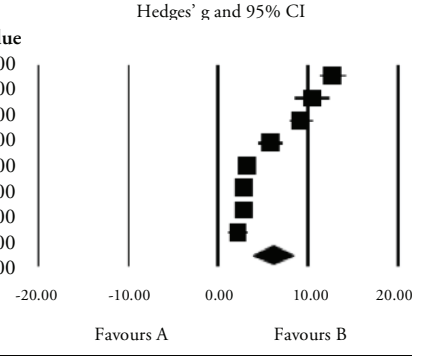

Figure 2. Meta-Analysis Table (High to Low Arrangement)

two levels were not different times, but one time, or were not caused by anything known, they would be discussed in the report and reported as one factor, not two, in a random design.

The Final Model

The final model was composed of two factors: 1) the eight different studies, and 2) the two time intervals. This was a random effects model for the various studies included, which had subgrouped times (1 minute and 30 seconds) embedded in them. The times were fixed effects. This provided a "mixed effects" model. The model selected was also 1) an analysis across levels of the two subgroups, and 2) a comparison of the effects of these subgroups.[10] This study had a common variance that was pooled. Figure 3 presents these data.

There were two sub-analyses occurring in this table. The first (summarized by the first diamond) was for a one-minute application, which provided a Hedges' *g* summary of 10.784 ($p < 0.000$). This was highly significant. There was also a 30-second application. This was not as effective as the one-minute application, but it, too, was very effective. It is summarized by the second diamond, a Hedges' *g* summary statistic was 3.317 ($p < 0.000$). The 30-second and one-minute applications were combined into an overall grand total Hedges' *g* statistic (the third diamond), which was 5.607 ($p < 0.000$).

Notice that there was still heterogeneity within these two times at the 0.05 level. Studies 13 and 26 were different from each other for the 1-minute application. Study 45 was different from all the other 30-second evaluations. They were not compared for homogeneity, because there was no indication they were handled differently. It was assumed that there was much variability among the studies. Therefore, they remained in the random effects model.

To formally test the 30-second and 1-minute application times, examine the two 95% confidence levels: 30 seconds = 2.378 – 4.258, and 1 minute = 9.372 – 12.197. The 30-second and 1-minute confidence intervals did not overlap, so they were different.

In summary, for the hand disinfectant data, the results indicated the product was effective at 30 seconds, but it killed many more bacteria when applied for one minute.

Looking for Bias in the Study

If the studies were all-inclusive in this analysis, then there would be no need to look for bias; however, it was unknown whether this occurred. There were two opportunities for bias to occur. The first, already discussed to some degree, was that studies opposed to the researcher's beliefs were eliminated. For example, a researcher may have chosen only the studies that showed their product superior to others.[3] To this end, significant studies were evaluated, and insignificant ones were not included in the evaluation. This is a major problem in meta-analysis.[1,2,3] The second case was that contradictory studies may not have been published. For example, very small studies or studies showing no effects

META ANALYSIS

Group by Subgroup within study		Study Name	Subgroup within study	Hedges' g	Standard Error	Variance	Statistics for each study: Lower limit	Upper limit	z-value	p-value
1 min		12.000	1 min	10.491	0.991	0.982	8.549	12.434	10.586	0.000
1 min	1 min	13.000	1 min	9.211	0.681	0.464	7.876	10.545	13.528	0.000
1 min		26.000	1 min	12.702	0.751	0.564	11.229	14.174	16.910	0.000
1 min	Total(1 min)		30 sec	10.784	0.721	0.519	9.372	12.197	14.965	0.000
30 sec		24.000	30 sec	3.312	0.352	0.124	2.622	4.001	9.415	0.000
30 sec		45.000	30 sec	5.773	0.716	0.513	4.369	7.176	8.063	0.000
30 sec	30 sec	48.000	30 sec	2.837	0.495	0.245	1.868	3.807	5.737	0.000
30 sec		67.000	30 sec	2.213	0.553	0.306	1.129	3.297	4.002	0.000
30 sec		71.000	30 sec	2.926	0.258	0.068	2.421	3.431	11.356	0.000
30 sec	Total (30 sec)			3.317	0.479	0.230	2.378	4.258	6.922	0.000
Overall	**Grand Total**			5.607	0.399	0.159	4.824	6.389	14.050	0.000

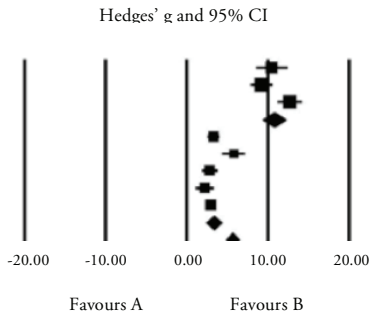

Figure 3. Meta-Analysis: Two Time Points (30 second and 1 minute)

are rarely published.[10] Studies that demonstrate large differences are more likely to be published than studies that do not. Both of these situations represent a potential bias in the meta-analysis.

Because bias cannot be avoided with certainty, its potential is assessed by formulating a few questions:

1. Is there evidence of bias?
2. Is it possible that the entire main effect is due to an artifact of bias?
3. How much impact of bias is present?

We simply do not know these answers, yet. The Cochrane Collaboration[6] has published the results of over 3700 meta-analyses and is a good place for the researcher to begin. It did not have any studies relevant to evaluating topical antimicrobials the way the FDA expects them to be evaluated. There were several studies listed that compared the incidence of disease relative to hand-washing, but this did not coincide with the design of this meta-analysis.

A good place to look for bias in this study was with a funnel plot,[16] which appears as funnel-shaped, or a graph composed of standard error versus the Hedges' *g* statistic (Figure 4). Generally, the smaller the study, the larger the standard error. The standard error is larger in a small study because the values in the numerator are divided by a smaller number in the denominator. This is opposed to a larger *n*, for larger sample sizes in a study, which give a smaller standard error. The funnel shape is caused by ordering the standard errors of the study. For this analysis, the smaller studies with larger standard error were plotted in the bottom portion of the graph; the larger studies with smaller standard error in the top portion (Figure 4).

It is similar to Exploratory Data Analysis (EDA), which, in general statistics, examines the data distribution.[13,14] Looking at a stem-leaf display, a researcher can see if the data fit a normal distribution.[14] Non-biased data would look like Figure 5A. If the data were biased; however, it could be seen that the lower values were removed (Figure 5B). The distribution looks abnormal on the stem-leaf display.

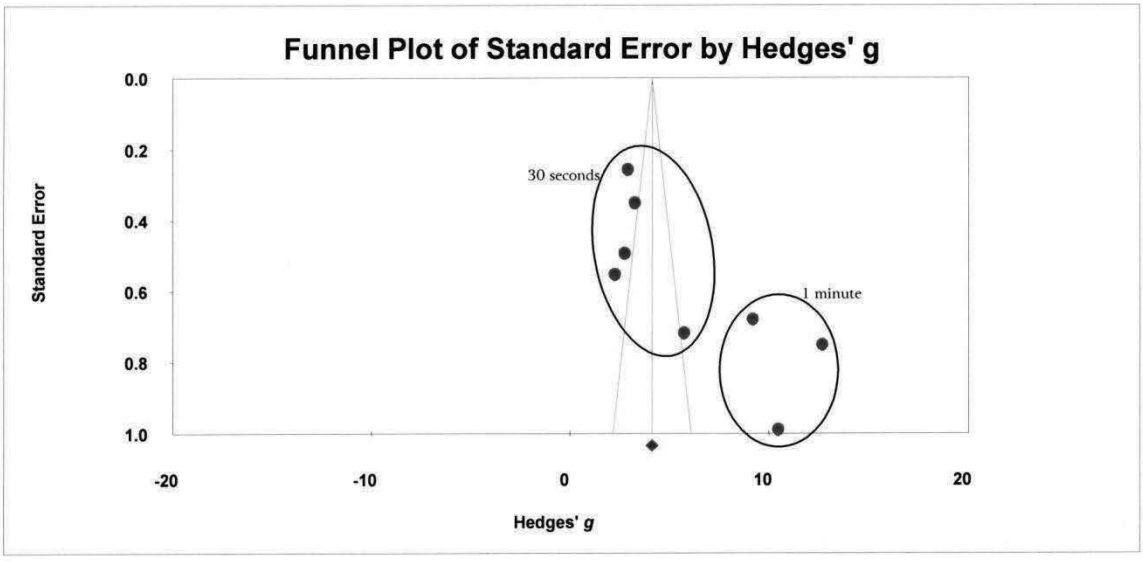

Figure 4. Funnel Plot

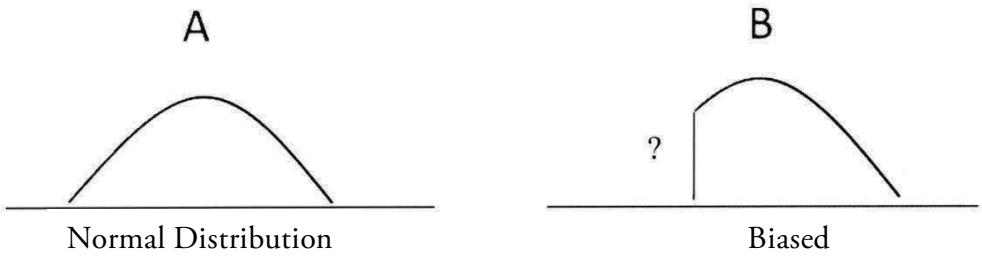

Figure 5. Stem-Leaf Distributions

In meta-analysis, the funnel plot serves a similar condition. In this study, however, there was a problem: two different time frames. The 30-second and 1-minute times were initially separated and two different funnel plots generated. However, the studies were limited to only three data points for the 1-minute time and only five for the 30-second time. This was not enough data to detect a bias if one existed, so the study remained undivided.

It was apparent that the three highest Hedges' *g* studies were performed at one-minute application times instead of 30-second application times. As a result, they pulled the Hedges' *g* to the right (to a higher Hedges' *g*). Another statistic, Duval and Tweedie's Trim and Fill Statistic,[17] used an iterative procedure to remove the most extreme studies by presenting mirror image of the most extreme data points in the graph (Figure 6). Using the Duval and Tweedie's trim and fill statistic showed the mirror image of the two most weighted studies as neutralized. It essentially canceled the two studies with high values by presenting a mirror image of them on the graph. Figure 6 shows the effects.

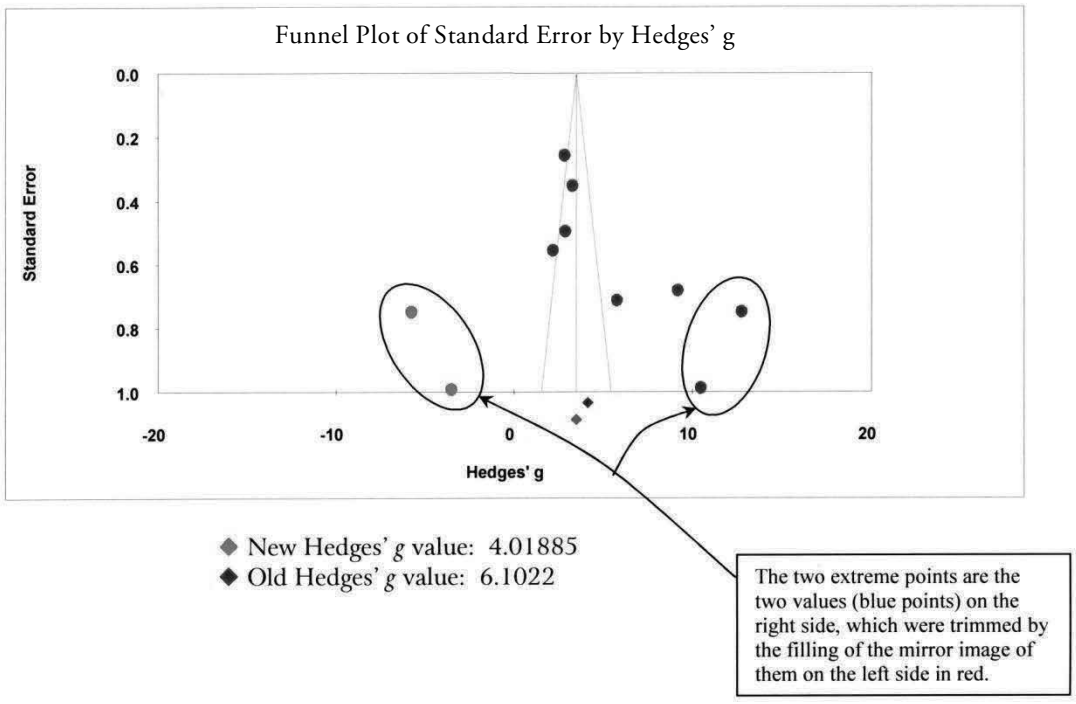

Figure 6. Funnel Plot of Standard Error by Hedges' *g* Using the Duval and Tweedie's Trim and Fill Statistic.

The actual value of the point estimate and the 95% confidence interval are shown in Table 2. The average grand total point went from 6.10220 to 4.01885. The upper and lower confidence intervals were also moved to the left (from 3.91971 – 8.28468 to 1.56302 – 6.47467) but were still significant, because zero was not included in the confidence interval.

Table 2. Duval and Tweedie's Trim and Fill

			Random Effects	
	Studies Trimmed	**Point Estimate**	**Lower Limit**	**Upper Limit**
Observed values		6.10220	3.91971	8.28468
Adjusted Values	2	4.01885	1.56302	6.47467

Note that this is a misrepresentation of this study, for three studies were performed at one-minute application times and five were performed at 30 seconds. At worst, the study results continue to be significant, even though the average effect has moved to the left. The three questions were then addressed. It was not known if bias was present, but there was no evidence of bias. It was possible that the main effect was due to bias, but if there was any, it was inconsequential; the product was still significant.

Conclusion

A researcher should present the data relative to the readers' comprehension, and remember that most readers are not statisticians. The key questions will be "what is easier for readers to understand?" and "How can data best be presented to them?" Meta-analysis allows one to integrate the results of various studies to achieve comprehensive understanding of the studies performed.

REFERENCES

1. Cooper H, Hedges LV, Valentine JG. The Handbook of Research Synthesis and Meta-Analysis. 2nd Ed. New York: Russell Sage Foundation; 2009.
2. Egger M, Smith GD, Altman DB. Systematic Reviews in Healthcare: Meta-Analysis in Context. London: BMA House; 2009.
3. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. West Sussex: Wiley & Sons; 2009.
4. Federal Register. 21 Code of Federal Regulations, parts 333 and 369. Tentative Final Monograph for Health Care Antiseptic Drug Products. 1994;31:402-52.
5. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. New York: Wiley – Blackwell; 2008.
6. The online Cochrane reports (Cochrane Collaboration). http://www.cochrane.org
7. Rothstein HR, Sutton AJ, Borenstein M. Publication Bias in Meta-Analysis. West Sussex: Wiley & Sons; 2005.
8. www.Meta-Analysis.com (This is the site from which this program was obtained. I used this program in writing this paper; it is a very useful system and easy to use.)
9. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effects and random-effects models for meta-analysis. Res Syn Meth 2010;97-111.
10. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones RR. Empirical assessment of effect of publication bias on meta-analysis. Biostat Meth J 2000;320:1574-7.
11. Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. London: Academic Press; 1985.
12. Hedges L. Distribution theory for Glass' estimator of effect size and relational estimations. Journal of Educational Statistics 1981;6:107-28.
13. Paulson DS. Applied statistical designs for the researcher. New York: Marcel Dekker, 2003.
14. Velkman PF, Hoaglin DG. Applications, basis, and computing Exploratory Data Analysis (EDA). Boston, MA: Duxbury Press, 1981.
15. Paulson DS. Biostatistics and microbiology. New York: Springer, 2008.
16. Light RJ, Singer JD, Willet JB. The visual presentation and interpretation of meta-analysis. In H Cooper & LV Hedges (eds.), The Handbook of Research Synthesis. New York: Russel Sage foundation, 1994.
17. Duval S, Tweedie R. Trim and fill: A simple funnel plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics, 2000;56:455-63.