2.1.67 Meta-analysis: Overview

Abstract

The term "meta-analysis", coined in the 1970's, describes the quantitative review of a number of different studies of the same phenomenon. The specific goals of meta-analysis include the estimation of an overall effect, across the different studies; the use of multiple studies to enable a more powerful test of the statistical significance of the effect; and identification of covariates affecting the estimated effect in different studies. Amongst the difficulties in using meta-analysis are problems of heterogeneity, due to combining unlike studies; and systematic problems due to"publication bias" or poor quality of studies.

1 What is meta-analysis?

1.1 The origins of meta-analysis

Meta-analysis (sometimes called "quantitative synthesis" or "overview analysis") is the term used to describe *quantitative* methods for combining information across different studies. The term "meta-analysis" seems to have been coined by Glass (1976) to describe this idea of utilizing information in many studies of the same effect, although the concept itself is very much older (dating back at least to the 1930's, when it was studied by Fisher and Pearson).

Glass (1976) also introduced the idea of combining different summary statistics from different studies in a scale-free form (known as "effect sizes"). Most commonly, in the sociological literature these forms include standardized mean differences and correlation coefficients. These techniques extend the applicability of the concept of meta-analysis, since one then does not need identical measures of the effect in each of the studies considered.

The ideas have proved to be very powerful, and since their introduction there has been a veritable explosion in the use of such techniques, with the prime growth probably occurring in analysis of sociological effects and medical or epidemiological results.

Of course, it has long been common to see reviews of an area, with an expert bringing together the different information and synthesizing a conclusion from many disparate sources. Such overviews often contain qualitative or subjective impressions of the totality of information available. In this sense, the idea of combining study information is an old and appealing one, especially when considering subtle effects that might be hard to assess conclusively in one study. The key contributions of meta-analysis lie in various attempts to *formalize* this approach, and the term is usually reserved for the situation where one is combining numerical effect sizes from a collection of studies, rather than giving a more general non-quantitative overview.

1.2 The goals of meta-analysis

Meta-analysis has become particularly popular in situations where the studies individually do not show that an effect is statistically significant. In this context, it is often the case that a combination of studies is more powerful in evaluating the effect. Methods can be divided into several groups:

(i) those which enable the *overall significance of an effect* to be evaluated, based on the multiple studies available;

(ii) those which attempt to estimate an *overall effect size* θ, by combining the individual estimates in multiple studies in appropriate ways;

(iii) those which evaluate the *heterogeneity* in a group of studies, often as a prelude to carrying out (i) and (ii);

(iv) those which evaluate possible *systematic biases* in meta-analytic methods.

We will give a brief overview of ideas in each of these groups, and illustrate them with an analysis of one specific example.

It must be stressed that underlying the general popularity of meta-analysis is the assumption that for most effects of interest, there is some grand overall value of θ that can actually be found more accurately by summarizing the information in many studies. If this concept is correct, then the methods are of considerable power and value.

However, in reality the value of the estimated effect size in a particular study is conditional on many specific and non-generalizable constraints pertaining in that study. To generalize, or even to combine, the values from such studies is not necessarily valid unless these constraints are reasonably satisfied across the range to which generalization is sought.

One option to deal with this is to develop linear or hierarchical models which incorporate covariate information (Hedges and Olkin 1985; DuMouchel 1991), or to use a response-surface approach as advocated in Rubin (1990). These methods often have the practical drawback that the information available on covariates in the individual studies is often sparse or non-existent, and that many studies do no more than announce that the estimates available are "adjusted" for some named or even un-named covariates.

Because of this partial knowledge of individual studies, it seems inevitable that many users will opt for the simple methods of meta-analysis described below, and will then take the overall values as applying much more widely than is often justified. At the very least, such naive approaches should be tempered by a careful evaluation of the heterogeneity involved.

1.3 IQ assessment: a typical meta-analysis

To carry out an ideal meta-analysis, there are a number of steps. We first identify an effect we wish to study. We then collect all of the studies on the subject, and combine their results using

a method of meta-analyis. This then gives us both an overall measure of the relationship, and a statistical assessment of the significance of the relationship taking into account all the studies.

As an illustrative example we will consider a set of 19 randomized studies of the effects of teacher expectancy on later pupil performance on an IQ test, taken from Raudenbusch and Bryk (1985) and analyzed also in detail in various chapters of Cooper and Hedges (1994). In each study the "treated" group consisted of students identified to their teachers as "likely to experience substantial intellectual growth", and the "control" group was not so identified. In this case, the effect size for each study represents the mean IQ score of the treated group minus the mean of the control group, divided by a pooled standard deviation: that is, each effect size in this example is a standardized mean difference.

The data are illustrated in a "ladder plot" in Figure 1. This plot is typical of that used in meta-analyses. Each of the studies is represented with its mean effect size together with the associated 95% confidence interval (CI). The size of the means in this illustration is proportional to the precision, indicating which studies are likely to carry more weight.

Figure 1 near here

As Figure 1 shows, the various effect sizes do not give a conclusive picture. In some studies the effect sizes are positive, in others negative. Only three are statistically significant on the face of it. The goal of meta-analysis is to try to combine these in some effective way.

2 A formal overview of meta-analysis

2.1 Testing significance of an effect

In the typical formalism for meta-analysis, we assume that study i provides a value T_i (the effect size), all of which are assumed to measure the same overall effect θ; and we also assume we know the standard error ν_i of the i^{th} effect size. One of the non-trivial aspects of a meta-analysis is the collection of these T_i, and care must be taken to extract appropriate information. It is easy to bias the outcome of a meta-analysis by careless selection of the T_i, and methods such as blinding of the extractors of the effect sizes, or multiple extraction by independent reviewers, are often used.

The first goal of meta-analysis is to try and decide whether or not an overall effect is significant. There are two common and simple approaches to this. The first is just vote-counting: how many studies have positive and how many have negative effect sizes? If there is no overall effect this should be a binomial variable. In the IQ example of Figure 1, such vote-counting yields 11 out of 19 in favor of a positive effect, but clearly this is not significant ($p = 0.33$).

This simple minded approach clearly fails to take into account any precision attached to each study. A somewhat more sophisticated idea, of "combining p-values", was introduced by Fisher (1932). Here the individual p-value is taken to encapsulate the information in the individual study, and all other aspects of the study are ignored in combining this information. If p_i is the p-value of the i^{th} study, since $\chi = -2\sum_1^n \log p_i$ has a χ^2_{2n} distribution under the null hypothesis, χ can be used to

assess whether the totality of the p-values leads to rejection or not. In the IQ example, this value is around 70 on 38 d.f., indicating now that the null hypothesis is rejected at the 99% level. It is worth noting that rounding in reporting of the effect sizes and variances can lead to surprising inaccuracies in the evaluation of χ; and that the significance here is very largely due to just two or three significant studies, with the negative and neutral studies having limited ability to overcome these strong positive results.

2.2 Combining effect sizes: fixed effects

While the combined p-value approach is valid under the appropriate assumptions of independence between studies, it has the clear drawback that it does not permit any estimate of an overall effect. Other approaches to meta-analysis use somewhat more information from the individual studies, and attempt to combine them in a sharper way. In the simplest such method (the so-called "fixed effects" model), the effect size is assumed to be normally distributed, so that formally, for the i^{th} study we thus assume

$$T_i \quad = \quad \theta + e_i; \tag{1}$$

it is assumed that e_i are independent $N(0, \nu_i^2)$ random variables. To use this type of approach some steps may be needed to render the assumption of normality at least approximately correct: for example, if the effect size is a correlation, a transform to normality is required. In other cases effect sizes may be (possibly standardized) differences between means, as in the IQ example; in yet other cases, they may be the estimates of slopes in a regression model, and so on.

The goal is now is to estimate the overall value θ, assuming the homogeneity inherent in (1). Typically it is assumed that the standard errors ν_i are known accurately, in which case standard theory indicates that θ is best estimated by

$$\overline{T}_\bullet = \left[\sum_i T_i/\nu_i^2\right] / \left[\sum_i 1/\nu_i^2\right]$$

which has variance

$$\sigma_\bullet^2 = \left[\sum_i 1/\nu_i^2\right]^{-1}.$$

In the IQ data, this fixed effects model leads to an overall estimate of the difference between means of $\overline{T}_\bullet = 0.06$, with a 95% CI of (-0.01,0.13). Although most of the individual studies are not significant, we have thus been able to use meta-analysis to establish that, at least on the face of it, the overall effect is significant at the 10% level, though not at the 5% level: not a strong result but one indicating that further study might well be warranted.

2.3 Combining effect sizes: random effects and non-homogeneity

The fixed effects model does not allow for heterogeneity between studies. When there is an indication that the studies are not homogeneous, it is common to combine estimates via a "random

effects" model (Draper *et al* 1993), which attempts to allow for inter-study variation. In the random effects model we consider the formalization

$$\begin{aligned} T_i &= \theta_i + \varepsilon_i \\ \theta_i &= \mu_\bullet + \xi_i; \end{aligned} \tag{2}$$

here T_i is the observed effect size for each study, θ_i is the corresponding true i^{th} effect size, and it is assumed that ε_i are independent $N(0, \nu_i^2)$ random variables, that the ξ_i are independent $N(0, \tau^2)$ random variables, and that the xi_i and ε_i are mutually independent. The fixed effects model takes $\tau^2 = 0$; by allowing $\tau^2 > 0$, the random effects model enables us to capture some of the inhomogeneity since it assumes different studies have mean values θ_i which may differ from $\mu_\bullet$.

In this case the meta-analysis estimatorof $\mu_\bullet$ is given by

$$\overline{T}_\bullet^* = \left[\sum_i T_i/[\nu_i^2 + \widehat{\tau^2}]\right] / \left[\sum_i 1/[\nu_i^2 + \widehat{\tau^2}]\right]$$

which has variance

$$\sigma_\bullet^{2*} = \left[\sum_i 1/[\nu_i^2 + \widehat{\tau^2}]\right]^{-1}.$$

There are various methods to give the estimator $\widehat{\tau^2}$, the most common of which is the estimator of DerSimonian-Laird (1986), whose variance is given in Biggerstaff and Tweedie (1997).

The random effects model can also be analyzed in a Bayesian context, and extends logically to hierarchical models (DuMouchel 1990; Draper *et al* 1993), by the addition of priors on θ and τ^2.

In the IQ data, this random effects model leads to an overall estimate of the difference between means of overall estimate of $\overline{T}_\bullet^* = 0.089$, with 95% CI $(-0.020, 0.199)$. The DerSimonian-Laird estimator of $\tau^2 = 0.026$, with a 95% CI of (0.004, 0.095), indicating significant lack of homogeneity; and now we see that by allowing for this heterogeneity, the significance of the overall $\overline{T}_\bullet^*$ is at 11%.

This of course indicates a different conclusion from, say, the method of combining p-values. The difference is explained by the rather different rejection regions underlying the different methods, and in general the results from combining effect sizes will be preferred, as they use considerably more detailed information.

2.4 Combining effect sizes: using covariates

One further extension of (2) is to incorporate covariates, in the form

$$\theta_i = \beta_0 + \beta_1 X_i + \varepsilon_i; \tag{3}$$

where X_i is a vector of covariates in study i and β_1 is a vector of parameters.

This is attractive when the individual studies contain sufficient information to enable the model to be fitted, since it helps explain the variability in the random effects model.

In our IQ example, there exist data on the length of time (in weeks) that the teachers are exposed to the children being tested. When this is factored into the model, it is found that the estimate of $\beta_0 = 0.424$ and the estimate of $\beta_1 = -0.168$, and both are highly significant (see Chapter 20 of Cooper and Hedges (1994)). In this case the covariate appears to explain much of what is going on: in all but one of the negative results, the teacher had known the children for more than 2 weeks, but only in one of the positive studies was this the case. Thus without direct knowledge of childrens' abilities, there seems to be a real effect of the treatment; but direct knowledge mitigates this almost entirely.

3 Problems with meta-analyses

3.1 Possible difficulties

There are several provisos that need to be taken into account before accepting a formal summation of the studies as in the section above, and with the huge increase in the use of meta-analysis, there has come a large number of books and discussion papers which assess the benefits, drawbacks and problems of these techniques (Glass *et al* 1981; Hedges and Olkin 1985; Draper *et al* 1993; Cooper and Hedges 1994; Mengersen *et al* 1995). Three of the key concerns that meta-analysis raises, and which differ from those in general statistical methodology, are:

(i) the problem of comparability of data and study design, since for the meta-analysis to be meaningfully interpreted, we must not combine "apples and oranges";

(ii) the effect of "publication bias", recognizing that failure to obtain all relevant studies, both published and unpublished, may result in a quite distorted meta-analysis.

(iii) the effect of different quality in different studies, so that one should not rely totally evenly on the studies used.

3.2 Are we comparing apples with oranges?

Meta-analysis is designed to enable combination of results from studies which are comparable. The interpretation of comparability is a subjective and often difficult one. In order to paint an honest picture of the aims and applicability of any meta-analysis, we must first carefully define the relevant effect with which we are concerned, and ensure that all studies collected do address this same effect. This can be quite non-trivial.

In the IQ example, for example, we would need to be sure that the tests for IQ were measuring similar attributes. Some comparability (at least of scaling of the test) is provided by the standardization of the mean differences. We also need to be convinced that the concept of "teacher expectancy" that we are evaluating is appropriately similar across the different studies, and from the written papers this is not always easy to decide.

There are three different ways one might suggest for handling such heterogeneity. The first is by using models that specifically allow for such an effect, such as the random effects models above. More subtly, to allow for the types of inhomogeneity we are concerned with, Bayesian methods might well be used. In this context the priors on the various parameters can perhaps be thought of, not as describing "prior information" in any strong sense, but rather as describing in more detail the way in which the studies might be heterogeneous.

A second method of handling variability is by building more complex models where covariates are introduced. This is clearly preferable when it can explain the variability otherwise swept into τ^2 in the random effects model. There are some who advocate that the random effects model should never be used, but that one should rather search out appropriate covariates to explain heterogeneity between studies, and as we have seen in the IQ example, this can be very fruitful. The drawback to this is that, since meta-analysis seeks to use the *published* results of studies without recourse to raw data which is often lost or unavailable, the user is often unable to use covariates since these are not published in sufficient detail.

A third (very simple) method used to account for heterogeneity is to give results separately for different subsets of the data which are thought to be heterogeneous, rather than to attempt to develop a parametric model for the effects of this stratification. This also is only possible if there are sufficient studies to allow reasonable estimation in each stratum.

3.3 Publication bias

One of the most interesting phenomena in meta-analysis is "publication bias".

It is obviously important in principle in meta-analysis to attempt to collect all published and unpublished studies relevant to the relationship in question. The problem here is that unpublished studies, by their nature, are likely to differ from published studies. They are likely to be less significant, since journals differentially accept significant studies. They are likely to show less "interesting" effects, since such studies are often not submitted; or in the case of non-English speaking authors, are submitted only to local journals that are missed in scanning. Hence their omission from a meta-analysis may well bias the combined result away from the null value.

Missing studies due to publication bias are not easy to allow for. Unlike traditional missing data problems, there is an unknown number of them. Their effect could be huge, or it could be minute, and developing a sensitivity analysis that accounts for them is not trivial. Publication bias seems to be a new form of bias that needs new methods to account for it.

There are several ways used to evaluate the existence and effect of such bias. The first is the "funnel plot", introduced in Light and Pillemer (1984), which gives a good graphical indication of the possible existence of some forms of publication bias. If one plots the effect size against some measure of size of study, then under the normal assumptions of the fixed and random effects models, there should be symmetry around the true θ; and since (for practical reasons) there are generally more small studies than large ones, one should typically see a funnel or tree shape for the pattern

of data. If the plot does not exhibit such symmetry then one might deduce that there are missing studies. This is illustrated on the IQ data in Figure 2.

Such graphical indications are the most frequently used diagnostic for publication bias, but give little information on what difference the "missing studies" might make. There are a number of rather complex approaches to this problem (Iyengar and Greenhouse 1988; Berlin *et al* 1989; Dear and Begg 1992; Hedges 1992; Givens *et al* 1997). In Duval and Tweedie (2000) a simpler method for handling such studies is developed which seems to give results consistent with more complex methods and quantifies the subjective impression given by using funnel plots.

FIGURE 2 NEAR HERE

For the IQ data, the methods in Duval and Tweedie (2000) estimate that the number of missing studies is around 2-3, with positions as indicated in Figure 2. Allowing for three such missing studies leads to a random effects estimate of $\overline{T}^*_\bullet$ of 0.027 with 95% CI of (-0.10, 0.16): that is, much of the observed estimate of θ might well be due to studies not included in the collection. Such a sensitivity analysis can aid in assuring that we do not become overconfident in assuming we have a full and correct estimate of the final answer.

3.4 Quality of studies

Clearly different studies are of different quality, and there is considerable debate about whether to exclude studies that are unreliable.

A policy of deliberate exclusion of poor quality studies also helps in many cases to mitigate the problems of publication bias. If the studies that are not published are poor quality, which is quite conceivable, then there may be reasons for excluding them even if they exist on the fringes of research publication.

Some quality aspects are readily agreed on. For example, there is general concensus that studies which are randomized in some way are better than purely observational studies. In the medical literature, the Cochrane Collaboration, which is attempting to develop a full set of information on various diseases and treatments, will only accept studies which are randomized clinical trials into its base of studies for inclusion. However, while there may be a rationale for only using (or conducting) randomized trials, in many sociological areas there is little possibility of using other than observational trials, and so this objective criterion for inclusion is not always of use.

There has been some work done on methods of allowing for quality (Cooper and Hedges 1994). Most of these methods involve weighting schemes, where the weighted averages in (1) are modified to depend, not just on the variances of the studies, but on other attributes of the studies. One such approach consists of drawing up lists of quality "attributes" and then, based on a formal scoring of papers, to weight according to the quality.

The problem with most schemes for assessing and accounting for quality differences is their subjectivity. In general, it seems best to ensure that studies are included rather than having some

excluded or down-weighted on grounds which are not clear and open. If there are real concerns about the quality of studies, then a viable alternative is to construct the analysis with and without these studies: as with many areas of meta-analysis, such sensitivity considerations can rapidly settle the role of the poor quality studies in the overall outcome.

4 Implementing meta-analyses

4.1 Collecting Data

There is no formal way of ensuring that all sources of data have been covered in developing the values for a meta-analysis. However, most meta-analyses at least attempt to carry out searches of all relevant data-bases, and then work from this list to a wider search. In the IQ example there are various sources of literature (relevant to many other sociological and educational meta-analyses) that might be formally searched: for example, the ERIC (Educational Resources Informational Center) database, *PsycINFO* and *Psychological Abstracts*, or *Sociological Abstracts*. The list will vary from field to field.

As well as such formal and usually computerized searches, it is also valuable to use other more informal methods. Following up on references in articles already found (especially review articles), consideration of citation indexes, and general conversations and communications with others in the field, will all assist in locating studies. In particular the last form, informal followup, is perhaps the best method for finding the otherwise missing studies, unpublished theses, or non-mainstream articles whose omission may lead to publication bias.

In all cases, it is imperative that the meta-analyst gives a full and detailed description of the search process used and the actual data derived. This is of particular importance in situations where the basic article gives more than one summary statistic that might be used.

4.2 Software for calculations

In order to apply the ideas above it would be ideal to point the potential user to appropriate software that could carry out the full range of meta-analysis. The ideal software is not yet available, although there are many homegrown versions in the statistical literature, with a variety of features, not all of them intuitively easy and (rather more problematically) not all of them giving correct results.

Methods to use SAS or BUGS to carry out both frequentist and Bayesian meta-analyses are described in Normand (1999) and DuMouchel and Normand (2000), and there is a range of recent SAS macros described in Wang and Bushman (1999). The Cochrane Collaboration, which aims to become a full registry of studies in clinical trials areas, also has developed some analytic software although this has to date only been available for studies in their collection. Various commercial software packages are currently under development which have many of the desirable features required by the non-expert.

Nonetheless, there is still a long way to go before meta-analysis can be carried out totally routinely.

4.3 Conclusions

The IQ example with which this overview is illustrated indicates many of the advantages and some of the pitfalls of implementing a meta-analysis.

The advantages are three-fold. We have been able to establish that, despite the existence of positive and negative studies, the overall effect is positive. We have found that, when lack of homogeneity is taken into account, the positive effect is not yet known to be statistically significant. And we have seen that the influence of covariates in these data sets may well be crucial, so that when they are taken into account, a much more clearcut picture appears.

The pitfalls are several. We have seen that the simplistic use of voting procedures, combined p-values or fixed effects models may give conflicting answers, and much thought needs to go into deciding how to use random effects or possibly Bayesian models in these circumstances. We have indicated that on the face of it, there may well be publication bias in this data set, and that this might account for much of the observed overall effect.

The implementation of this series of meta-analyses used a number of one-off pieces of software, for analysis and for graphical presentation. As this example shows, however, even when the mathematical methodology becomes routine to implement, there will still be a need for the practitioner to take every precaution to ensure that the results really do reflect a coherent picture of the overall effect being evaluated.

5 Further Reading

Detailed reviews of almost all aspects of meta-analysis are given in the books by Glass *et al* (1981), Hedges and Olkin (1985), Rosenthal (1991) and Cooper and Hedges (1994). A very readable account of the general problems of combining information is in Light and Pillemer (1984).

For related entries see META-ANALYSIS TOOLS and META-ANALYSIS IN PRACTICE, and for more sophisticated models of the same type see HIERARCHICAL LINEAR MODELS.

References

Berlin J A , Begg C B , Louis T A 1989 An assessment of publication bias using a sample of published clinical trials. *J. Amer. Statist. Assoc.*, 84:381–392

Biggerstaff B J Tweedie R L 1997 Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine*, 16:753–768

Cooper H, Hedges L V (eds.) 1994 *The Handbook of Research Synthesis.* Russell Sage Foundation, New York, N.Y.

Dear KBG, Begg CB 1992 An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*, 7:237–245

DerSimonian R, Laird N M 1986 Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7:177–188

DuMouchel WH, Normand S-L T 2000 Computer modeling and graphical strategies for meta-analysis. In Stangl DK, Berry DA (eds) *Meta-analysis in Medicine and Health Policy*, pages 127–178. Marcel Dekker, N.Y..

DuMouchel WH 1990 Bayesian meta-analysis. In D. Berry, editor, *Statistical Methods for Pharmacology*, pages 509–529. Marcel Dekker, N.Y.

Draper D, Gaver DP, Goel PK, Greenhouse JB, Hedges LV, Morris CN, Tucker JR, Waterman C 1993 *Combining Information: Statistical Issues and Opportunities for Research.* American Statistical Association, Washington

Duval SJ, Tweedie RL 2000 A non-parametric "trim and fill" method of assessing publication bias in meta-analysis. *J. Amer. Statist. Assoc.*, 95:89–98

Fisher RA 1932 *Statistical Methods for Research Workers.* Oliver and Boyd, Fourth edition.

Givens GH, Smith DD, Tweedie RL 1997 Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate (with discussion). *Statistical Science*, 12:221–250

Glass GV 1976 Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5:3–8

Glass GV, MgGaw B, Smith ML 1981 *Meta-analysis of Social Research.* Sage Publications

Hedges LV 1992 Modeling publication selection effects in meta-analysis. *Statistical Science*, 7:246–255

Hedges LV, Olkin, I *Statistical Methods for Meta-analysis.* Academic Press

Iyengar S, Greenhouse, JB 1988 Selection models and the file drawer problem. *Statistical Science*, 3:109–135

Light RJ, Pillemer DB 1984 *Summing Up: the Science of Reviewing Research.* Harvard Univ. Press

Mengersen KL, Tweedie RL, Biggerstaff, BJ 1995 The impact of method choice in meta-analysis. *Australian J. Statistics*, 37:19–44

Normand S-L T 1999 Meta-analysis: Formulating, evaluating, combining and reporting. *Statistics in Medicine*, 18:321–359

Raudenbush, SW, Bryk, AS 19885 Empirical Bayes meta-analysis. *Journal of Educational Statistics*, 106:75–98

Rosenthal R 1991 *Meta-analysis Procedures for Social Research*, Sage Publications, Newbury Park.

Rubin D 1990 A new perspective. In Wachter, K, Straf M (eds), *The Future of Meta-analysis*, pages 155–166. Russell Sage Foundation, N.Y.

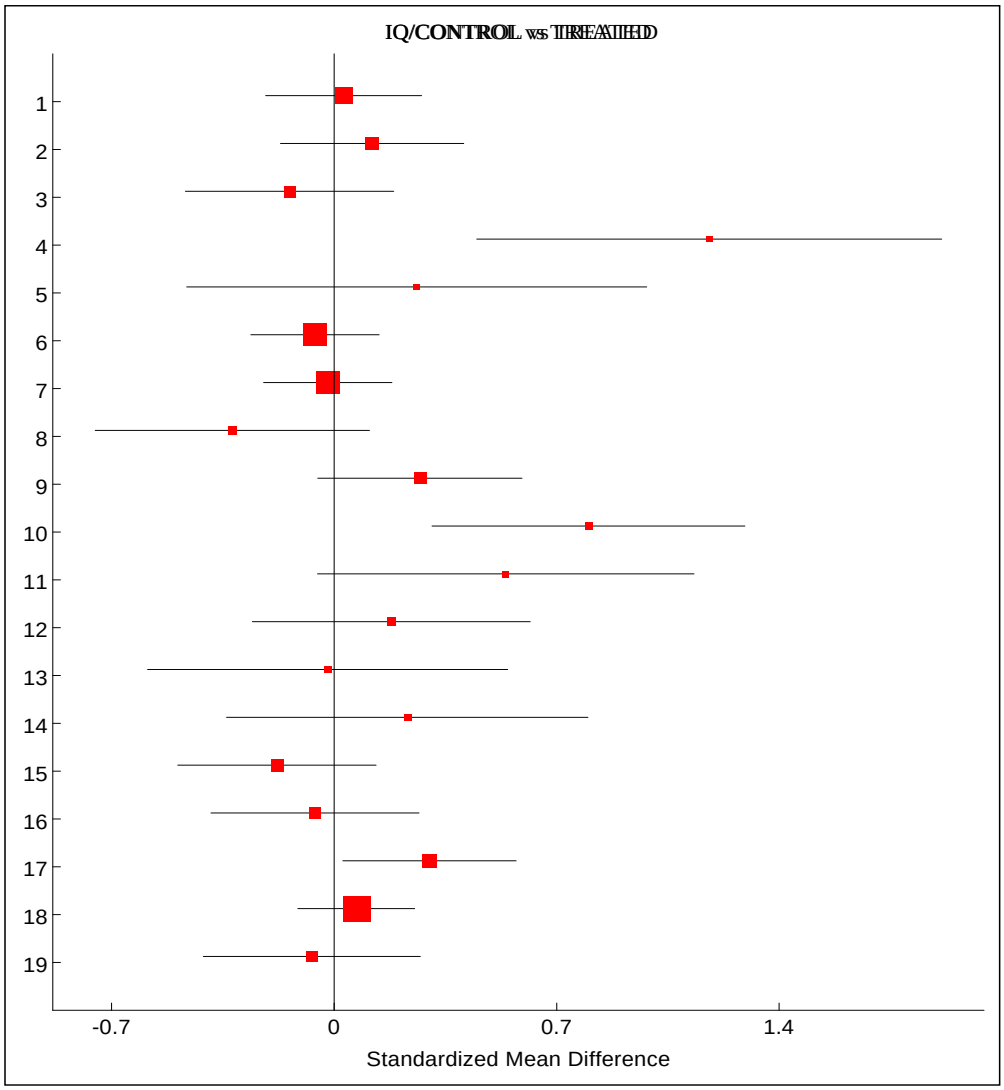
Wang MC, Bushman BJ 1999 *Integrating Results through Meta-analytic Review Using SAS Software.* SAS Institute.

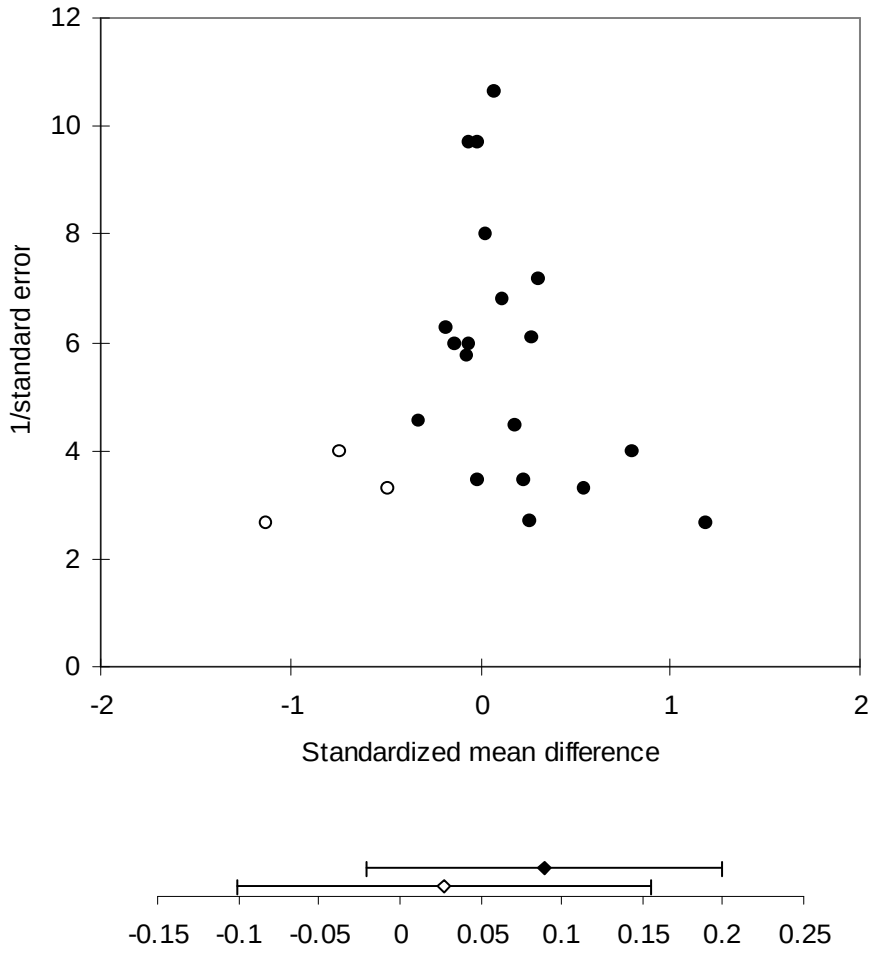
R.L. Tweedie
Division of Biostatistics
School of Public Health
University of Minnesota
Minneapolis, MN 55455, USA

Figure Titles

Figure 1: Ladder plot of 19 studies of student IQ modified by teacher expectations. Size of squares is proportional to accuracy of study

Figure 2: Possible publication bias in studies of teacher expectancy of IQ. Top panel is a funnel plot of standardized mean differences: the solid circles are original data, the open circles are three imputed "missing studies". Bottom panel shows overall mean and 95% CI before and after allowing for the missing studies.


IQ/CONTROL vs TREATED
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
-0.7
0
0.7
1.4
Standardized Mean Difference


12
10
8
6
4
2
0
1/standard error
-2
-1
0
1
2
Standardized mean difference
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25