Gee, T. (2005) Capturing study influence: The concept of 'gravity' in meta-analysis, *Counselling, Psychotherapy, and Health*, 1(1), 52-75, July 2005.

Capturing study influence: The concept of 'gravity' in meta-analysis

Dr. Travis Gee

School of Human Services
Griffith University
Meadowbrook, Qld.

Centre of National Research on Disability
and Rehabilitation Medicine

Mayne School of Medicine

University of Queensland
Herston, Qld., Australia

Address correspondence to Dr Travis Gee, at gee@ccs.uq.edu.au

Gee, T. (2005) Capturing study influence: The concept of 'gravity' in meta-analysis, *Counselling, Psychotherapy, and Health*, 1(1), 52-75, July 2005.

Abstract

While much theory abounds over properties of meta-analyses, there appears to be very little work to date on examining jackknifed and exact distributions of the statistics generated by the method. This paper takes an initial step towards filling that gap by describing a SAS (SAS Institute, 2001) macro written by the author based on another macro due to others, that performs jackknifed estimates of average effect size. It also suggests that 'gravity' is a property of studies included in a meta-analysis. A computer simulation supports the utility of the approach, and proposals for future development of exact and approximate methods in 'combinatorial meta analysis' are set forward.

Acknowledgement:
Thanks to Joan Hendrikz of CONROD at the UQ School of Medicine for comments on the final draft of this article.

Meta-analysis (MA) is a method that uses statistical techniques to aggregate the results of other studies (Glass 1976; Glass, McGaw and Smith 1981); Rosenthal, 1991; Rosenthal and DiMatteo, 2001). It is promoted perhaps most strongly by the Cochrane Collaboration, which aims to use the method to substantiate the evidence base for all medical practice (see http://www.cochrane.org) and which provides reports on studies such as that of Montgomery and Dennis (2003), which combined the results of six trials that comprised 282 participants, and using meta-analytic procedures demonstrated that there is a case to be made for the use of cognitive-behavioral therapy for sleep problems in older adults.

Critical commentary has been sparse, but a growing chorus suggests that interpretation of meta-analyses is not as straightforward as some would assume (Ernst and Pittler, 2001; Gatchel and McGeary, 2001; Gee, Bellamy & Campbell, 2005; Hopayian, 2001; Olsen, Middleton, Ezzo, et al., 2001; Shrier, 2003). The present article adds to that chorus by suggesting that the current asymptotic theory that underlies detection of outliers in meta-analysis may be supplanted by more exact methods that do not rest upon distributional assumptions.

“Effect sizes” are the crux of the matter in MA, as these are the things that get added up to find out whether the effect of a treatment is large (or the relationship between two variables is strong). These are most clearly conceptualized in terms of a standard normal distribution, with mean zero and standard deviation 1 (denoted $N[0,1]$). For the purposes of this article, we are considering the case where we would be comparing two groups; say, those who received a treatment versus those who received a placebo, or who were in a wait-list control group, then an effect size (ES) of 1.0 is noted if the mean of the treatment group was one standard deviation away from the mean of the other group. This implies that the average member of the treatment group fared better than 84% of the other group. For this reason, an effect size of 1.0 is considered large. Smaller effect sizes are more common, and may be interpreted roughly with reference to the rule that .2 is small (average treatment did better than 54% of controls), .5 is moderate (treatment did better than 69% of controls) and .8 is the lower range of “large” (treatment did better than 79% of controls).

Expressing the ES as a standardized difference allows comparisons of studies that use different measures to quantify the same construct. It also allows the incorporation of research results that reflect other cases, such as the strength of association between measures instead of differences between groups. Pearson’s r, for example, can also be expressed in terms that allow comparison with studies that report t-tests. However, a great strength of meta-analysis is that all can be reduced to a common metric, the ES, which makes all such cases amenable to the method proposed in this paper. An existing piece of software, *meta.sas* (Dimakos & Friendly, 1997), exists for SAS users to compute meta-analytic statistics and is available on the Internet.

Needless to say, not all studies are constructed equally. A clinical sample might have twenty cases in each group, whereas a population survey might have a thousand. It is possible to estimate the effect size both with and without reference to the number of studies. Common meta-analysis routines provide both weighted and unweighted means that reflect the overall ES for a given set of studies. Unweighted means are not commonly taken as serious indicators of overall ES because in any given MA, studies can range widely in terms of sample size. The weighted mean difference (WMD)

between two groups is a common statistic, which is usually taken as a good indicator of effect size, and can be used when studies all use the same measure. However, where multiple measures must be combined, the standardised mean difference, or SMD must be used (see http://www.cochrane-net.org/openlearning/HTML/modA1-4.htm for a discussion). Whilst there are numerous formulae by which study statistics can be converted to such indicators, for the present purposes, the traditional difference-between-two-means divided by pooled-standard-deviation is taken as the indicator. However, as noted above, any other measure can be subjected to the method proposed below.

Proposing Gravity as a Property of Studies in a Meta-Analysis

As to weighting, consider that if two otherwise-equivalent studies provided ES estimates of .2 and .4, then the unweighted estimate of the average effect size would be 0.3, which is equal to ES= (.2+.4) / 2 . However, if the first study had a sample size of 1000, and the second a sample size of 50, then the weighting procedures usually employed would result in the average being rather close to 0.2, because in the formula, the weight assigned to the study with ES=0.2 would be a good deal higher than the weight assigned to the study where ES=0.4.

Weighting is important, as it allows for the fact that due to the Central Limit Theorem, small studies can produce larger ES estimates just due to chance fluctuations in sampling. Such deviations would have undue influence unless we accounted for variation in *n*. The analogy to make here is that studies with a large sample size may influence the ES estimated by a meta-analytic procedure, and cause it to 'gravitate' towards the value of the 'weightier' study. Thus, we may consider 'gravity' as a property of studies in a meta-analysis, such that "negative gravity" would be present in studies which, when removed, cause the ES estimate to drop, and "positive gravity" would be present in studies which, when removed, cause the ES estimate to increase.

Typically, meta-analyses comprise a good deal more than two studies, but for simplicity, let us initially consider the scenario where there are three effect size estimates from three studies: .1, .2, and .9. Clearly the last is inconsistent with the other two, and the mean, .4 is substantially above the median, .2. Here, an "outlier" has created some "pull" away from what would (if the last study were excluded) be an average around .15 (at least, if the N's were equal in the two studies that produced the smaller ES values). If we assume equal N's in all studies, then the "pull" will be directly proportional to the distance of the outlying study from the average of the other two. However, if the first two studies had sample sizes of several hundred and the large-ES study had eight people in each group, random variation could possibly account for the large deviation in the smaller study's ES (and the estimated ES for the combination would certainly be closer to .2). In MA, the low N would result in less weight being placed on the small, deviant study, bringing it's 'gravitational' effect on the overall analysis down. The effect of the various studies on the overall average thus depends both on ES and N.

There are thus two sources of this "pull," to which I will henceforth refer as 'gravity.' Heterogeneity amongst studies is a problem that reflects variability in the extent to which some studies may unduly influence a meta-analysis. As pointed out by Sidik and Jonkman (2005) "Valid inference about an overall treatment effect in random-effects meta-analysis depends on accurately quantifying such heterogeneity among studies." It is of course important to understand the substantive differences that exist between studies in the literature (eg. different sample sizes, outcome measures, treatment qualities, study designs, etc.). However, it is also important to grasp the extent to which any single study may pull (or fail to pull) the overall ES estimate towards it. Key to the method presented herein is the idea of "jackknifing."

Jackknifing

The key statistical technique to understand for the present application is that of jackknifing. Proposed by Quenouille (1949) and developed by Tukey (1958), it is a technique that has only really come into its own since the development of the computer, because it requires a great deal of computation. It is now recognised in standard texts (eg., Efron & Tibshirani, 1993).

The idea is simple. The spread of values obtained as statistics can be examined by eliminating each observation from the dataset, which creates perturbations in the estimate. As a simple example, if we have the scores 1, 2 and 3, the average is 2. But if we eliminate 1, it becomes as high as 2.5. Eliminating 2 has no effect, and eliminating 3 drags it down to 1.5. Observations may be considered "outliers" (see Kruskal's classic 1960 paper) when the effect that removing them has on the statistics that are computed becomes disproportionately large.

The technique is important, because whenever a statistic is estimated, there is some degree of error associated with it. In complex situations, the distribution of an average, for instance, might not conform to a simple normal bell curve. If there are a dozen means taken from a dozen studies that have a dozen different sample sizes, then the expected sampling distribution of those means is a composite of 12 variance estimates weighted by 12 sample sizes. The assumption that normality will make it all well and good is tenuous at best, and the suggestion by Lanyon (1987) that "Jackknifing and bootstrapping should be used in all cases where a statistic is generated and the distribution for that statistic is unknown or too complicated for the more conventional methods of dispersion estimation" rings true for MA, which almost invariably involves such complexities.

This method affords a direct approach to the estimation of the impact of a study on a meta-analysis, as it examines the effect of removing it directly, thereby permitting estimation of the degree of perturbation associated with the study relative to the remaining ones. Given how long it has been established as a method, it is surprising that a jackknifing approach to studying the relative contributions to studies in MA has not been taken before, however, to the present writer's knowledge, it does not appear to have been considered outside of the its use in metaregression, where results are modelled on the basis of predicting ES from various methodological features of a study.

Approaching Gravity

The present approach to meta-analytic gravitation is based upon the premise that elimination of a study from a meta-analysis will tend to change the estimate of effect size to some degree, but not in the same way as is achieved with jackknifing. This is due to the fact that the 'pull' of a study on MA results is achieved not only through ES alone, but through the weighting due to sample size. Jackknifing, however, is a tool that can be used to study the phenomenon. A 'massive' study with high gravity but relatively low ES will, when removed, cause the ES estimate to be higher than it would be when that study is included. Equally, a 'massive' study with high gravity but relatively high ES will, when removed, cause the ES estimate to drop. Removal of low-gravity studies will have but little effect, but the *distribution* of the relative magnitudes in the set of studies may provide a yardstick by which a study might be identified as a potential outlier.

Measuring Gravity

Given the preceding considerations, it is possible to examine several features of the ES estimates. It is necessary to apply some shorthand to the terms we require. We will term the overall ES estimate that is obtained for all available studies $ES_{o,}$ the average jackknifed effect estimate. If we denote the effect size that is obtained on the run where study *i* is excluded as ES_i,

$$ESo = \sum \left(\frac{ESi}{k}\right) \qquad 1.$$

If each ES_i is an estimate of ES_j, the deviations of ES_i around ES_j should distribute normally and thus the dispersion of these estimates may be described as the variance of the ES_i,

$$Var\,(ESj) = \sum (ESi - ESj)^2/(k-1) \qquad 2.$$

which is the squared standard deviation of the individual estimates

$$Sj = \left(\sum (ESi - ESj)^2/(k-1)\right)^{\left(\frac{1}{2}\right)} \qquad 3.$$

for *k* studies. This implies a normal distribution of the perturbation-based ES estimates, which in turn means that it is possible to refer to a *Z* table for estimates of probability. However, if the studies have a heterogeneous mixture such that some come from a population where the true ES is different than for the rest, it should be possible to identify such studies as outliers.

Study 1: Behaviour of Gravity in Homogeneous Samples

It is useful and instructive to study properties of populations of studies such as those typically seen in meta-analysis with computer simulation methods. The first property so considered here is a certain horseshoe effect that is implied by the fact that gravity will be affected both by sample size and effect size. To the extent that studies approximate the average ES their gravity will be small, almost irrespective of sample

size. However, as sample size grows, smaller deviations from the average will be expected to carry more weight, and gravity will be sizable even at values closer to ES_j. Thus, a certain "horseshoe" shape should emerge from large groups of studies that are subjected to jackknife estimation of gravity. This will be less visible according to the degree of homogeneity of the studies in terms of both ES and sample size.

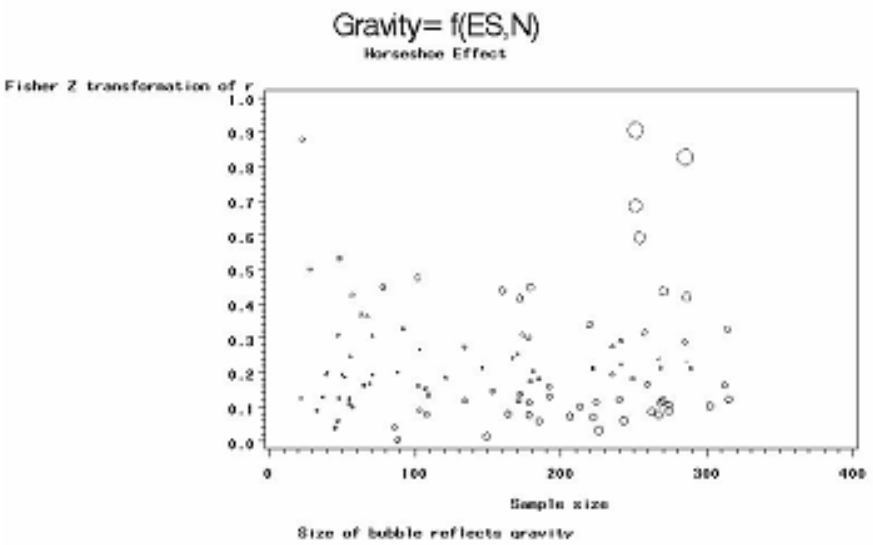
Method

To study this, 100 simulated study results were generated using the SAS System (SAS Institute) with a fixed modelled effect size of .3. The two-group t-test statistic was computed on simulated continuous data, with sample sizes modelled using a random number between 10 and 320 for the control group N, and a random number within +/- 0.3 of the control group N for the treatment group N. Standard deviations were modelled on a chi-square distribution (using n random Z^2 values, dividing by *df* = n-1), and taking the square root for each group.

These data points then were subjected to jackknifing as described above. Estimates of gravity were obtained using the SAS macro *jackmeta.sas* (see Appendix A), and are plotted in Fig. 1.

Results

Fig. 1. Gravitational Horseshoe



Sample size (N) and ES (using the Fisher r to Z transformation applied in *meta.sas*) are plotted in Fig. 1, with circles showing the relative gravity of each study. The horseshoe shape created by the high-gravity (large circle) studies is plainly visible.

Discussion – Study 1

The visibile horseshoe effect in Fig. 1 is consistent with the predicted behaviour of gravity as defined above. However, we must consider that the present results were modelled based on a control mean of zero and a fixed treatment effect of .3, with unit standard deviation. To study the horseshoe effect, variability in ES was not modelled. What is also visible, though, is a small group of rather extreme values that combine large sample sizes with large effect sizes. The extent to which such studies might have an undue influence raises the question of outliers in such studies. Fortunately, the concept of 'gravity' provides some access to whether or not a study may be considered sufficiently deviant to exclude it from further analysis, and therefore interpret it separately in the overall meta-analysis.

Study 2: Outlier Detection

We assumed before that gravity would be normally distributed around ES_j with a standard deviation as described in Eq. 2. On this assumption, we can compute *z* and *p*-values for the perturbation associated with the removal of each of *k* individual studies. This may of course be adjusted as the user wishes, as in the *jackmeta.sas* macro there is an option where alpha for outlier detection can be set to 5%, 1% or whatever value the user desires. This theory was tested using computer simulation, and a nominal alpha value of 5%.

Method

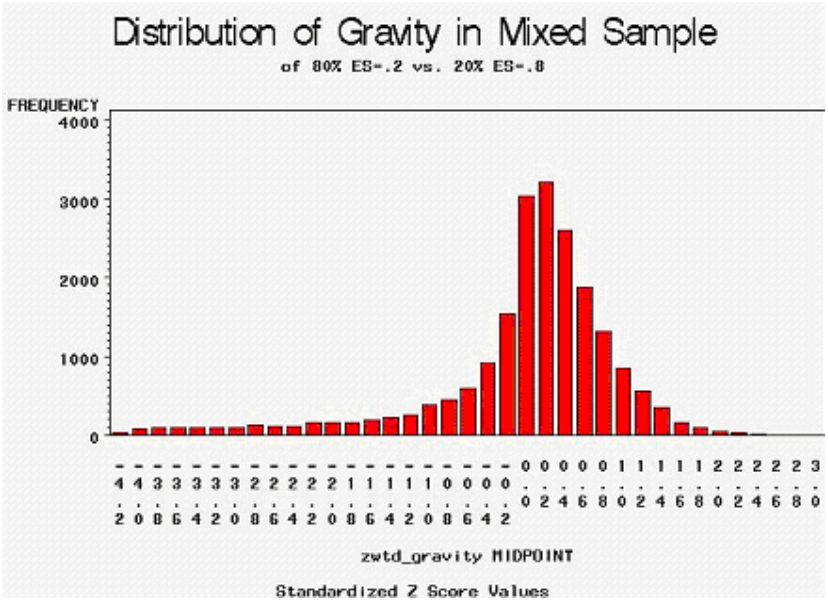
To examine the utility of using gravity to identify single outliers in sets of studies, 1000 simulations were run of sets of 20 studies, where the effect size in each simulation was set to 0.2 (small) for 5 studies, and 0.8 (large) for one study, the "outlier." The two-group t-test statistic was computed on simulated continuous data, with sample sizes modelled using a random number between 10 and 160 for the control group N, and a random number within +/-3 of the control group N for the treatment group N. Standard deviations were modelled on a chi-square distribution (using n random Z^2 values, dividing by *df* = n-1 and taking the square root for each group). The *jackmeta.sas* macro was run on each set of simulated values. Studies were identified as 'outliers' if the perturbation associated with removal of that study created a deviation from the overall result that had a standardized score in excess of +/- 1.96 (for alpha=.05).

Results

Using alpha=.05 to identify cases as 'outliers,' 17842 out of 19000 'non-outlier' studies (94%) were correctly identified as non-outliers, and 493 of the 1000 'outlier' studies were correctly identified as such, meaning that just under half of true outliers were detected, as against the 5% that would be expected by chance methods. There were 1158 false positives (making the true positive rate 89%) and 507 false negatives. In signal detection terms, this means that sensitivity was .493, and specificity was .939.

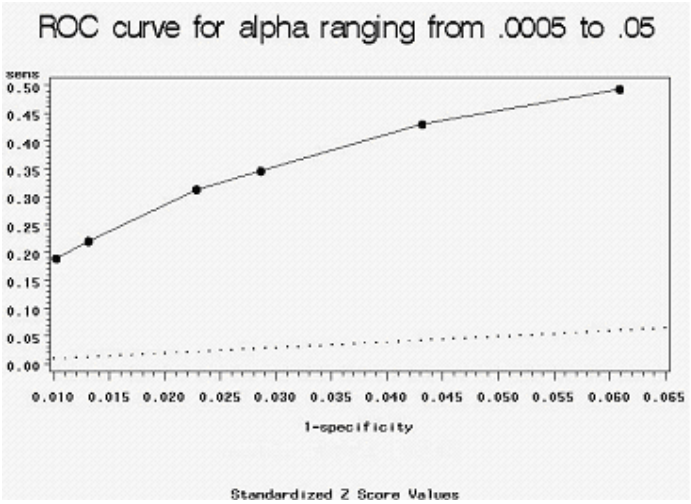
The distribution of the Z-scores for gravity estimates is provided in Fig. 2. The clear negative skew indicates that many more studies caused large drops in ES when omitted from the jackknife procedures than would be expected by chance, even though only 1 study in 20 (5%) were modelled to be 'outliers.'

Fig. 2 Distribution of Standardized Gravity Estimates



Re-examining the data using alpha =.01 led to detection of 34.6% of genuine outliers, and 97% of genuine non-outliers (sensitivity=.346, specificity=.971). Alpha=.005 led to 31.3% detection of true outliers, and 97.7% of true non-outliers (sensitivity=.313, specificity=.977). The portion of the ROC curve for the alpha range .0005 to .05 is presented in Fig. 3.

Fig. 3 ROC curve for range of alpha cutoff values (dotted line for reference)



Discussion – Study 2

The concept of gravity applied in jackknifed analysis of even modest sets of twenty studies has shown, through simulation, that deviant effect sizes that differ by .5 are detectible through the use of gravity when a single outlier is present in the dataset. However, more complex situations are likely to exist, which suggests that extending the study of gravity to study sets that may contain high-gravity studies that may oppose one another, for instance, is necessary. To accomplish this will require a generalisation of the jackknife method as described below.

Conclusions and Future Directions: More Complex Cases

A conceptual basis for a programme of research into jackknife and other developments combinatorial methods in meta-analysis has been put forward, with some demonstration of its utility in establishing 'gravitational' properties of studies, which has promise to become a useful tool in understanding the effects of studies on MA. The present simulation suggests that jackknifing can detect outliers in modest sets of data using the gravity method, and identifying as outliers those cases where effects of either sample size or effect size (or some combination thereof) may produce undue influence on a result. Refinement of the method is needed however, to reduce the false positive and false negative rates that are apparent at least in the present simulation.

The limitations to this are of course that more study is needed to identify the extent to which such deviations may be detectable in more complex datasets, and to examine the accuracy of prediction across ranges of the variables that were modelled (eg., sample size patterns, magnitude of deviation of outliers from the balance of studies, etc.). The jackknifing approach and generalisations of it may offer a door into these and other elements as well, as studies exist in a literature that has its own properties. For example, jackknifed estimates may behave differently, taking on particular distributions, when a literature contains many small studies and few large ones, as against literatures that contain many population-based surveys and few small-scale studies. Further modelling of such patterns is possible with the *jackmeta.sas* macro.

Another limitation is that models for binary responses have not been considered, nor for correlation coefficient data. The behaviour of these models under jackknifing conditions is an area that needs exploration. As these have been written into the original *meta.sas* macro, this research is in a position to proceed rapidly. However, there have been other ways of estimating effect size put forward by other writers (eg., Deeks, 1999) which are not yet coded into the macro, but which could easily be incorporated with some programming, so that alternative methods can be compared in terms of their behaviour under combinatorial conditions. For instance, are clusters of low-gravity studies less likely to emerge when using Hedges' adjusted *g* (see Deeks, 1999), which corrects for small sample size?

The present demonstration was of course limited in scope, as this is an introductory article aimed only to point towards new directions in meta-analysis. It is not, for example, always the case that there is a single outlying study in a set of studies subjected to MA. Indeed, it does not make sense to assume that this would be the case. The more modern random-intercepts model of MA assumes that a variety of treatment effects may be present, particularly in studies with multiple followup times,

and sets of studies where hetereogeneous followups were used to measure effects (eg., some followup at three weeks, others at 10). However, there are a couple of generalisations that are obvious based on the preceding development, which bear mentioning here as they are currently being researched by the present author.

Future Direction #1: Exact Combinatorial Meta-Analysis

In view of the preceding, it makes sense that this simple illustration of the case of a single outlier be followed up in future research with a method that allows gravitational effects to be studied when random effects may be present, and outliers of varying sizes and directions may affect the results. To extend the idea, consider that in the present method, we may regard the jackknifed estimates to be a case of k studies taken k-1 at a time. This is a special case, and we are not limited to that approach. With the power of modern computers, it is possible for small-to-modest-sized sets of studies to run what might be termed a 'combinatorial meta-analysis,' which is the set of all possible meta-analyses of k studies taken r at a time, for r=1 to k. The behaviour of all possible subsets of studies can then provide a background against which to identify subgroups of studies of size r against other size-r clusters that may or may not share common features.

For example, pairs of high-gravity studies that are opposite in sign would be expected to retain relatively high-gravity properties when analysed together, while pairs of high-gravity studies that are similar in sign and magnitude will decrease markedly in overall gravity when studied together, as both will approximate their mean ES with a good deal of 'force' that is attributable to the sample size component of gravity. Identification of local minima and maxima is therefore theoretically a possible basis on which clustering of studies could be performed, thereby 'quantifying sources of heterogeneity,' as pointed out by Sidik and Jonkman (2005).

Such a method also helps resolve conceptual problems in the debates that often surround meta-analysis over whether some subset of studies should or should not be included. By computing all possible meta-analyses, the effect of inclusion/exclusion of certain studies or combinations of studies can be placed into a context, and indeed, the net gravity of the disputed subset can be computed directly and referred to a jackknifed normal distribution to test the significance of the argument that there is something indeed special about them that skews the results. "In/out" arguments that are but tempests in teapots can thus be placed into context, and where there is a full cyclone in the teapot, our meta-analytic meterology stands a better chance of detecting it against a background of relative 'calm.'

Future Direction #2: Approximate Combinatorial Meta-Analysis

Naturally, there are computational limits to what can be performed. The foreseeable future of such combinatorial methods in meta-analysis must allow for the huge range of possible combinations when the number of studies available for analysis becomes large. When exact methods of computing indices become unwieldy (ie., when the number of combinations is prohibitively large), approximate methods may be used, in which random samples of the studies are meta-analysed, and the various properties of the combinations could be studied empirically. For example, where varying followup times are used across studies (eg., 1 week, 2 weeks or 3 weeks), the effect of the treatment at the different times could be considered as a function of the number of

studies using 1, 2 or 3 week followups that are included in a given MA. The obvious case is to examine the 1-week vs. the 2-week vs. the 3-week studies.

The more subtle approach, however, is to study how the 1-week group behaves as 'impurities' are added from the other two groups. Thus, where specific subsets of studies are at issue, these could be compared against a generated comparison distribution that excludes them, and again, gravitational effects of sub-clusters may be referred to a simulated distribution based on all other studies. This can be applied to differential follow-up times, groups of studies that share methodological characteristics, groups of studies that share certain measures but differ from other studies on that feature, and so forth.

Summary

Both the exact and approximate methods proposed to extend jackknifing seem, on the face of it, to be susceptible to Glass' criticism, that "the population [of studies] is nothing but the sample write large and we really know nothing more than what the sample tells us in spite of the fact that we have attached misleadingly precise probability numbers to the result" (Glass, 2000). Surely, we must keep in mind that p-values for simulation studies are meaningless, as runs can be made arbitrarily large. However, where we study the inner workings of a set of studies, we merely echo, in a systematic and unarguable way, that some studies can be 'in' and others 'out' with identifiable effects. Furthermore, by studying the *range* of combinations, we allow the relative merit of specific 'in/out' arguments to be evaluated against all others.

The proposed methods also afford researchers the opportunity to study the behaviour of research result sets that have particular characteristics. Literatures that have a large number of small studies, for instance, may plausibly differ from literatures that have a small number of large studies, and the problems may be different. The present method has the potential to allow researchers to explore how patterns of sample size may affect meta-analyses. Extensions of the method may allow estimation of dispersion of ES estimates, for instance, by permuting the sample size/effect size combinations to find exact distributions that emerge from those perturbations of the raw data.

In summary, jackknifing is a special case of a more general method which holds promise to extend the present findings to larger datasets with more complex properties. Software to perform exact and approximate combinatorial meta-analyses is currently being written to extend the present findings. However, even in the current state of development, it is clear that simple jackknifing provides, through the notion of 'gravity,' a window into the less-complicated cases of outliers that may exist in meta-analyses of small to moderate size.

References

Deeks J (1999). Statistical methods programmed in MetaView Version 4, paper on behalf of the Statistical Methods Working Group of the Cochrane Collaboration. Online document, http://www.a3.san.gva.es/mbe/statisticalmethods4.pdf , accessed 9 Mar. 2005.

Dimakos, I., and Friendly, M., (1997). *Meta.sas*. [software macro for the *SAS* System]. Online resource available at http://euclid.psych.yorku.ca/ftp/sas/macros/meta.sas.

Efron B, and Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton: Chapman & Hall/CRC, 1993, pp. 141–51.

Ernst E and Pittler MH, (2001). Assessment of therapeutic safety in systematic reviews: Literature review. *BMJ;323, September 8*:546.

Gatchel RJ, and McGeary D. Cochrane collaboration-based reviews of health-care interventions: are they unequivocal and valid scientifically, or simply nihilistic? *Spine 2001,26(2)*:196-205.

Gee, TL., Bellamy, N., and Campbell, J., (2005). *Bugs in the Black Box: A Birds' Eye View of Some Hard-To-Digest Aspects of Cochrane Reviews*. Unpublished m.s. in preparation for submission.

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

Glass, G., (2000). Meta-analysis at 25. Jan. 2000. Online resource at http://glass.ed.asu.edu/gene/papers/meta25.html. Accessed 22/2/2005.

Glass, Gene V., McGaw, Barry, and Smith, Mary Lee, (1981). Meta-analysis in social research. Beverly Hills; Sage.

Hopayian K, (2001). The need for caution in interpreting high quality systematic reviews. *BMJ;323(7314):*681-4.

Kruskal, J., (1960). Some Remarks on Wild Observations. *Technometrics, 2(1),* 1-3. [Available online at http://www.tufts.edu/~gdallal/out.htm accessed 1/3/2005.]

Montgomery P, and Dennis J. (2003). Cognitive behavioural interventions for sleep problems in adults aged 60+. The Cochrane Database of Systematic Reviews 2003, Issue 1. Art. No.: CD003161. DOI: 10.1002/14651858.CD003161. Abstract available online at [http://www.cochrane.org/cochrane/revabstr/AB003161.htm].

Olsen O, Middleton P, Ezzo J, Gotzsche PC, Hadhazy V, Herxheimer A, Kleijnen J, and McIntosh H. (2001). Quality of Cochrane reviews: assessment of sample from 1998. *BMJ 2001,323(7317):*829–832.

Quenouille, M. (1949). Approximate tests of correlation in time series. Journal of the Royal Statistical Society, Soc. Series B, 11, 18-84.

Rosenthal, R., (1991). Meta-analytic procedures for social research. Revised edition. Newbury Park, CA: Sage.

Rosenthal, R., and DiMatteo, M.R., (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Ann. Rev. Psychol., 52,*59-82.

SAS Institute (2001). *The SAS System for Windows, Release 8.02*. Cary, NC: SAS Institute.

Shrier I. (2003). Cochrane reviews: New blocks on the kids. *Br J Sports Med;37*:473-474.

Sidik, K., and Jonkman, J.N., (2005). Simple heterogeneity variance estimation for meta-analysis. *Appl. Statist. 54(2),*367-384.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. Annals of Mathematical Statistics, 29, 614.

APPENDIX A: JACKMETA.SAS

```
/*************************************************************/
/*  NAME: jackmeta.sas                                       */
/*  TITLE: Calculates meta-analytic indices                  */
/*         according to the Rosenthal & Rubin method         */
/*                                                           */
/* AUTHOR: Ioannis C. Dimakos                                */
/* ORIGINAL: 24SEP95 - Presented at SUGI 21, p938-942        */
/* MODIFIED BY:  Michael Friendly (macrified) 28 Oct 1997 18:05  */
/* Revised: 31 Jul 2002 16:04:53                             */
/*   - Fixed bug with select                                 */
/* Revised: 5 Mar 2005 by T. Gee to do jackknifing           */
/*************************************************************/
 /*=
=Description:

 The meta macro calculates several meta-analytic indices from
 summary statistics (F, t, r, z, p, chisq) for two or more
 studies, each testing one or more hypotheses.  For each hypothesis,
 the program calculates an equivalent value of a z-statistic,
 a correlation (r), and Fisher Z transformation of r (zf).  These
 are summarized by unweighted and weighted means to give overall
 results.

=Usage:

 Prepare a data set containing one observation for each study-
hypothesis
 to be included in the meta analysis.  The following information must
 be recorded:
                the sample size (n) per hypothesis,
                the type of statistic (F, t, r, z, p, chisq),
                the observed value of statistic,
                degrees of freedom (df),
                and p-value for each hypothesis in the meta-analysis.

==Parameters:

        data=_last_,   Name of the input data set
        id=,           Names of one or more variable(s) which
identify study
                       & hypothesis
        n=n,           Name of the variable giving the Sample size
for the
                       hypothesis test
        stat=stat,     Name of test statistic (f, t, r, z, p, chisq)
        value=value,   Name of the variable giving the value of the
test statistic
        df=df,         degrees of freedom for the test statistic
        p=p,           p-value
        out=metanal    Name of the output data set


=References:
   Rosenthal, R. (1991). Meta-analytic procedures for social
research.
           Newbury Park, CA: Sage.
   Mullen, B. (1989). Advanced BASIC meta-analysis. Hillsdale, NJ:
                Lawrence Erlbaum Associates.
```

```
=Example:
%include goptions;

data studies;
        length authors $12;
        input study hyp stat $ value n df p authors &$:;
        label n='Sample size'
                hyp   ='Hypothesis'
                stat  ='Statistic used'
                value ='value of statistic'
                df    ='degrees of freedom'
                p     ='p-value';
cards;
  1  1  t   3.13   20  18 .0001   Bar & Foo 86
  1  2  t   2.03   10   8  .078   Bar & Foo 86
  2  1  r    .60   10   8  .067   Foo & Bar 89
  3  1  z   3.14   23   .   .     Foo 90
;
option spool;
%meta(data=studies, id=study hyp);
=*/


%global numrecds;
%macro numobs ( _sasdsn_ );
 data _null_;
 set &_sasdsn_ point=nobs nobs=nobs;
 call symput( 'numrecds', put( nobs, best.) );
 stop ;
 run ;
%mend numobs;

%macro jackmeta(
        data=_last_,
        id=study_id,  /* Variable(s) which identify study &
hypothesis */
        n=n,          /* Sample size for the hypothesis test
*/
        stat=stat,    /* Name of test statistic (f, t, r, z, p,
chisq) */
        value=value,  /* Value of the test statistic
*/
        df=df,        /* degrees of freedom for the test statistic
*/
        p=p,          /* p-value */
        out=metanal,  /*output file, overwritten regularly*/
        display=N,    /*display results? Y to display*/
            alpha=1,      /*alpha for finding outliers*/
        studynum=studynum /*sequence of study in dataset*/
        );

*proc printto log='NUL';

/*create dataset to link studynum with id*/
data tolink; set &data; keep studynum &id; run;
proc sort; by studynum;run;

data &out;
        set &data;
 /* transform initial criteria to meta-analytic criteria.  Use z
```

```
  * for significance level and r (and Fisher's z) for  effect size
  */
&studynum=_n_;
label &studynum = 'sequence of study in dataset';

select; * (&stat);
 when (&stat='t')
  do;
   z=sqrt(&df*(log(1+(&value**2/&df))))*sqrt(1-(1/(2*&df)));
   r=sqrt(&value**2 /(&value**2 + &df));
        if &p=. then &p = 1 - probt(&value,&df);
  end;

 when (&stat in ('f', 'F'))
  do;
   z=sqrt(&df*(log(1+(&value/&df))))*sqrt(1-(1/(2*&df)));
   r=sqrt(&value/(&value+&df));
  end;

 when (&stat='chisq')
  do;
   z=sqrt(&value);
   r=sqrt(&value/&n);
        if &p=. then &p = 1 - probchi(&value,&df);
  end;

 when (&stat in ('z', 'Z'))
  do;
   z=&value;
   r=sqrt(&value**2/&n);
  end;

 when (&stat='r')
  do;
   t=(&value*sqrt(&n-2))/sqrt(1-&value**2);
   z=sqrt(&df*(log(1+(t**2/&df))))*sqrt(1-(1/(2*&df)));
   r=&value;
  end;

 otherwise
  do;
   z=abs(probit(&value));
   r=sqrt(z**2/&n);
  end;

 end;
 zf=.5*(log((1+r)/(1-r)));

 label
  &p='p-value'
  z='z-value'
  zf='Fisher Z transformation of r'
  r='Pearson r';

/*
 Calculate:
 1) product of Sample Size n (the weight) and Z-score,
 2) squared sample size (w=n**2), to be used
    in estimating the combined significance level.
 3) Weight for Diffuse Comparison of Effect Sizes.
*/
```

```
 nz=&n*z;
 w=&n**2;
 wzf=&n-3;
run;

 /* Calculations of Combinations of Effect Sizes and Significance
  * Levels.  Calculations of Diffuse Comparisons of E.Ss and S.Ls
  * Use separate PROC MEANS to calculate the various meta-analytic
  * indices.

 Step 1. Mean Effect Size Unweighted and
         Weighted By Sample Size
*/

proc means noprint data=&out;
   var zf;
   output out=combzf1 mean=meanzf1;
   run;

proc means noprint data=&out;
   var zf;
   weight &n;
   output out=combzf2 mean=meanzf2;
   run;

 /* Step 2. Calculate chi^2 for Diffuse Comparison of Effect Sizes.
  * Chi^2 has k-1 degrees of freedom.
  */

proc means css noprint data=&out;
   var zf;
   weight wzf;
   output out=diffzf css=csszf;
   run;

 /* Step 3. Combinations and Diffuse Comparisons of S.L Calculate
  * sums of N*Z and Squared Weights to be used for Combination of
  * S.L, chi^2(df=k-1) for Diffuse Comparison of S.L.
  */

proc means noprint data=&out;
   var nz w;
   output out=sigcomb sum=sumnz sumw;
   run;

proc means noprint data=&out;
   var z;
   output out=sigdiff css=cssz;
   run;

/*
 Step 4. Final Calculations for
                Combined Significance Level,
                Probability of Significance Level, and
                Probability of chi^2 for Diffuse Comparison
                of Effect Sizes.
*/
data final;
   merge combzf1 combzf2 diffzf sigcomb sigdiff;
   zcomb=sumnz/sqrt(sumw);
   probcomb=1-probnorm(zcomb);
```

```
   probz=1-probchi(cssz,_FREQ_-1);
   dfz=_FREQ_-1;
   probzf=1-probchi(csszf,_FREQ_-1);
   dfzf=_FREQ_-1;
   keep meanzf1 meanzf2 zcomb cssz csszf
        dfz dfzf probcomb probz probzf;
   label meanzf1='Mean Effect Size, Unweighted'
      meanzf2   ="Mean Effect Size, Weighted by &n"
      zcomb     ='Z, Combination of Significance Levels'
      probcomb  ='Probability for Z'
      cssz      ='x2, Diffuse Comparison of Sig. Levels'
      probz     ='Probability of x2'
      dfz       ='degrees of Freedom'
      csszf     ='x2, Diffuse Comparison of Effect Sizes'
      probzf    ='Probability of x2'
      dfzf      ='degrees of Freedom';
   run;


%if &display=Y %then %do;
 /*
  Presentation Step 1.
  Print Primary Statistics of Individual Studies
 */
 proc print data=&out label uniform;
  %if %length(&id) %then
      %str(id &id;) ;
  var &n &stat &value &df &p z r zf;
  title  'Meta-Analysis: Initial Statistics and Transformations';
 run;

 /*
  Presentation Step 2.
  Print Meta-Analytic Statistics obtained with SAS
 */
 proc print data=final label uniform noobs;
  var meanzf1 meanzf2 zcomb probcomb cssz
      dfz probz csszf dfzf probzf;
  title2 'Combinations and Diffuse Comparisons';
  title3 'of Effect Sizes and Significance Levels';
  footnote;
 run;

 /*
  Presentation Step 3.
  Chart of Effect Sizes.
  Use PROC CHART if PROC GCHART unsupported.
 */
 proc gchart data=&out;
  vbar zf /  midpoints=0 to 1 by .2 raxis=axis1;
  axis1 label=(a=90 r=0);
  title 'Frequency Distribution of Effect Sizes';
 run;

 /*
  Presentation Step 4.
  Plot Effect Sizes against Sample Sizes (aka the funnel plot)
  Use PROC PLOT if PROC GPLOT unsupported
 */

 proc gplot data=&out;
```

```
  plot &n*zf/ haxis = 0 to 1 by .1 hminor=1
              vaxis = axis1;
  axis1 label=(a=90 r=0);
  title 'Plot of Fisher Zf and Sample Size';
 run; quit
%end;

/*assign overall result to macro variable to carry forward*/

data final; set final;
call symput( 'overall_es_unw', put( meanzf1, best.) );
call symput( 'overall_es_wtd', put( meanzf2, best.) );
run;


/****end of main meta-analysis module****/


/****begin jackknifing modules****/

data jackresults; delete; *kill off any old results files;

data jackdata; set &out; run;*count number of studies;
%numobs(jackdata);

/*loop for j studies, excluding one at at time*/

%do j=1 %to &numrecds;
title "Jackknife run # &j";
data jack; set jackdata;*jackknife excluding study j;
 if &studynum ne &j then output; else delete;
run;
/*
 Step 1. Mean Effect Size Unweighted and
         Weighted By Sample Size
*/

proc means noprint data=jack;
   var zf;
   output out=combzf1 mean=meanzf1;
   run;

proc means noprint data=jack;
   var zf;
   weight &n;
   output out=combzf2 mean=meanzf2;
   run;

 /* Step 2. Calculate chi^2 for Diffuse Comparison of Effect Sizes.
  * Chi^2 has k-1 degrees of freedom.
  */

proc means css noprint data=jack;
   var zf;
   weight wzf;
   output out=diffzf css=csszf;
   run;

 /* Step 3. Combinations and Diffuse Comparisons of S.L Calculate
  * sums of N*Z and Squared Weights to be used for Combination of
  * S.L, chi^2(df=k-1) for Diffuse Comparison of S.L.
```

```
  */

proc means noprint data=jack;
   var nz w;
   output out=sigcomb sum=sumnz sumw;
   run;

proc means noprint data=jack;
   var z;
   output out=sigdiff css=cssz;
   run;

/*
 Step 4. Final Calculations for
                Combined Significance Level,
                Probability of Significance Level, and
                Probability of chi^2 for Diffuse Comparison
                of Effect Sizes.
*/
data final;
   merge combzf1 combzf2 diffzf sigcomb sigdiff;
   zcomb=sumnz/sqrt(sumw);
   probcomb=1-probnorm(zcomb);
   probz=1-probchi(cssz,_FREQ_-1);
   dfz=_FREQ_-1;
   probzf=1-probchi(csszf,_FREQ_-1);
   dfzf=_FREQ_-1;
   keep meanzf1 meanzf2 zcomb cssz csszf
        dfz dfzf probcomb probz probzf;
   label meanzf1='Mean Effect Size, Unweighted'
      meanzf2   ="Mean Effect Size, Weighted by &n"
      zcomb     ='Z, Combination of Significance Levels'
      probcomb  ='Probability for Z'
      cssz      ='x2, Diffuse Comparison of Sig. Levels'
      probz     ='Probability of x2'
      dfz       ='degrees of Freedom'
      csszf     ='x2, Diffuse Comparison of Effect Sizes'
      probzf    ='Probability of x2'
      dfzf      ='degrees of Freedom';
   run;

   data jackresults; set jackresults final(in=inb);
     if inb then studynum=&j;
   run;

%end;*end jackknifing loop;

/*get the overall jackknifed means & sds*/

title 'Overall Jackknife Results';

proc means noprint data=jackresults;
var meanzf1;
output out=jrzf1 mean=jack_mean_all_unw std=jack_std_all_unw;
run;

proc means noprint data=jackresults;
var meanzf2;
output out=jrzf2 mean=jack_mean_all_wtd std=jack_std_all_wtd;
run;
```

```
data jackstats; merge jrzf1 jrzf2;
label jack_mean_all_unw= 'ES estimate, unw.';
label jack_mean_all_wtd ='ES estimate, wtd.';
label jack_std_all_unw ='ES SD, unw.';
label jack_std_all_wtd ='ES SD, wtd.';
/*set values into macro variables*/
call symput( 'jack_mean_all_unw', put( jack_mean_all_unw, best.) );
call symput( 'jack_mean_all_wtd', put( jack_mean_all_wtd, best.) );
call symput( 'jack_std_all_unw', put( jack_std_all_unw, best.) );
call symput( 'jack_std_all_wtd', put( jack_std_all_wtd, best.) );
run;

/*link results with stats files*/

data jackresults; set jackresults;

/*assign values from macro to live variables*/
jack_mean_all_unw=&jack_mean_all_unw;
jack_mean_all_wtd=&jack_mean_all_wtd;
jack_std_all_unw=&jack_std_all_unw;
jack_std_all_wtd=&jack_std_all_wtd;
label jack_mean_all_unw= 'ES estimate, unw.';
label jack_mean_all_wtd ='ES estimate, wtd.';
label jack_std_all_unw ='ES SD, unw.';
label jack_std_all_wtd ='ES SD, wtd.';


/*compute gravity for unweighted results*/
unw_gravity=meanzf1-jack_mean_all_unw;
zunw_gravity=unw_gravity/jack_std_all_unw;

/*compute gravity for weighted results*/
wtd_gravity=meanzf2-jack_mean_all_wtd;
zwtd_gravity=wtd_gravity/jack_std_all_wtd;

/********look for outliers********/

/*Define zones of significance*/
%let siglevel=%sysevalf(&alpha/100);
%let siglower=%sysevalf(&siglevel/&numrecds);*Bonferroni correction;
%let sigupper=%sysevalf(1-&siglower);
%put siglower=&siglower sigupper=&sigupper;

pstd_grav_u=probnorm(zunw_gravity);
pstd_grav_w=probnorm(zwtd_gravity);
if &siglower < pstd_grav_u and &sigupper > pstd_grav_u then
unw_outlier=0; else unw_outlier=1;
if &siglower < pstd_grav_w and &sigupper > pstd_grav_u then
wtd_outlier=0; else wtd_outlier=1;
if %eval(&alpha/100) < pstd_grav_u <%eval(1-&alpha/100) then
unw_nobon_outlier=0; else unw_nobon_outlier=1;
if %eval(&alpha/100) < pstd_grav_w <%eval(1-&alpha/100) then
wtd_nobon_outlier=0; else wtd_nobon_outlier=1;
label zf='r->Z ES estimate';
alpha=%eval(&alpha/100);
drop _type_ _freq_;
run;

%if &display=Y %then %do;
  proc gchart; vbar unw_gravity -- pstd_grav_w;
  title 'Gravity Results';run;
```

```
%end;

*attach ID variables;
proc sort data=jackresults; by studynum;
data jackresults; merge METANAL jackresults tolink; by studynum;
run;

*PRINT RESULTS;
proc sort; by zwtd_gravity;run;

proc print noobs;
var &id zf wtd_gravity zwtd_gravity pstd_grav_w n diff treat_mean
ctrl_mean treat_sd ctrl_sd;
title 'Studies sorted by influence';
run;

PROC GPLOT; plot pstd_grav_w*(diff treat_mean ctrl_mean treat_sd
ctrl_sd);
run;

proc g3d; scatter n*zf=zwtd_gravity;run;
%mend;
proc gchart data=metanal; vbar zf;run;

proc printto log=LOG;

data hylangf20;
studynum=_n_;
length weeks $ 6;
length study_id $ 22;
input weeks $ Study_ID $ treat_tota treat_mean treat_sd
                         ctrl_total ctrl_mean ctrl_sd;
      stat="t";
      n=treat_tota+ctrl_total;
    df=n-2;
    diff=treat_mean-ctrl_mean;
    poolvar=((ctrl_total-1)*ctrl_sd**2+(treat_tota-
1)*treat_sd**2)/df;
    poolsd=sqrt(poolvar);
    se=poolsd*sqrt(1/ctrl_total+1/treat_tota);
    value=(treat_mean-ctrl_mean)/se;
    format p z7.6;
    p= 1-probt(value,df);
cards;
 1-4     Karlsson2002b(SvP)    86   45   25.67   33   44   30.39
 1-4     Moreland1993    46   47   27.13   48   51   27.71
 1-4     Scale1994a(2inj)    23   32   23.98   24   47   24.5
 1-4     Scale1994b(3_inj)    15   22   19.36   15   44   19.36
 1-4     Wobig1999c(NEhyl)    38   40   24.66   36   53   24
 1-4     Wobig1998    57   31   22.65   60   53   23.24
 5-13     Karlsson2002b(SvP)    86   41   31.53   33   46   34.9
 5-13     Scale1994a(2_inj)    23   27   23.98   24   53   24.5
 5-13     Scale1994b(3_inj)    15   11   19.36   15   43   19.36
 5-13     Wobig1999c(NEhyl)    37   32   24.33   35   43   23.66
 5-13     Wobig1998    57   23   22.65   60   60   23.24
 14-26     Karlsson2002b(SvP)    86   43   33.78   33   44   33.78
 14-26     Scale1994a(2_inj)    15   18   23.24   21   57   22.91
 14-26     Scale1994b(3_inj)    15   22   23.24   15   45   23.24
 14-26     Wobig1998    56   35   29.93   60   56   30.98
;
run;
```

```
%jackmeta(data=hylangf20,id=study_id weeks, display=Y);

proc varcomp data=jackresults method=reml;
class study_id weeks;
model zwtd_gravity=study_id weeks;
run;
proc gplot data=jackresults;
plot zunw_gravity*zwtd_gravity=study_id;run;

proc plot data=jackresults;
plot zwtd_gravity*zf $study_id;run;
```