15

Temporal Trends in Effect Sizes: Causes, Detection, and Implications

Julia Koricheva, Michael D. Jennions, and Joseph Lau

The general aim of meta-analysis, as well as of any other form of research synthesis, is to combine scientific evidence scattered through a number of individual studies addressing the same topic. Evidence, however, is not static and tends to evolve over time due to changes in research methods, changes in the characteristics of the subjects being studied, and so forth. New studies might either strengthen or challenge the conclusions of previous reports, resulting in changes in the mean effect size and its variance over time. The magnitude and direction of the mean effect size, and the breadth of its confidence interval, largely determine the conclusions drawn from a meta-analysis. Examples include whether or not a particular treatment, policy, or management strategy works; whether or not factor A has a biologically and/or statistically significant effect on the response variable X; or whether or not a hypothesis is supported by empirical tests. It is important therefore to be aware of the extent of temporal variation in effect sizes and to understand the reasons for this variation.

A number of recent studies in ecology and evolution (Table 15.1) have shown that temporal trends in effect sizes are common and often quite dramatic in these fields. This may perhaps reflect higher heterogeneity of studies included in ecological and evolutionary meta-analyses as compared to those in medicine (Chapter 25). Unlike other sources of heterogeneity, which affect the generality of conclusions drawn in meta-analysis, temporal changes in effect sizes might also jeopardize the stability of those conclusions (i.e., the conclusions of meta-analyses on the same topic conducted in different years might differ). Moreover, some of the methods used for detection of temporal trends in effect sizes, such as cumulative meta-analysis, differ from those used to detect other sources of heterogeneity. For the above reasons, we devote an entire chapter to temporal changes in effect sizes. We first summarize the findings of studies that examined temporal changes in the magnitude and direction of effect sizes in ecology, evolutionary biology, medicine, and the social sciences, and then discuss their possible causes, methods of detection, and implications for the interpretation of the results of the meta-analysis.

EVIDENCE OF TEMPORAL CHANGES IN EFFECT SIZES IN ECOLOGY AND EVOLUTIONARY BIOLOGY

Significant changes in the magnitude and even direction of research findings over time have been reported in many research syntheses from several different areas of ecology and evolutionary biology during the last two decades (Table 15.1). Notably, most of these areas represent

TABLE 15. 1. Outcomes of studies that have tested for temporal changes in reported effect sizes in ecology and evolutionary biology.

Reference	Topic	Pattern revealed
Alatalo et al. 1997	Heritability of secondary sexual characters	Approx. 2-fold increase in the reported estimates of heritability of male ornaments following publication of models supporting the assumptions of good-genes theory
Gontard-Danek & Møller 1999	Relationship between the strength of sexual selection and the expression of secondary sexual characters	Effect size was significantly negatively related to year of publication
Møller & Alatalo 1999	Relationship between male traits and offspring survival	Negative relationship between effect size and the year of publication
Simmons et al. 1999	Relationship between fluctuating asymmetry and sexual selection	Dramatic decrease in effect size and proportion of studies supporting the role of fluctuating asymmetry in sexual selection over <10 years
Poulin 2000	Effects of parasites on host behavior	Significant negative relationship between the effect size and the year of publication over the 30 years of research
Dubois & Cezilly 2002	Relationship between breeding success and mate retention in birds	Weak negative relationship between effect size and year of publication, which became nonsignificant once the effect of clutch size was controlled for
Gardner et al. 2003	Decline in Caribbean corals	Reduction in the rate of coral loss from 1980s to 1990s
Leimu and Koricheva 2004	(1) Effects of N fertilization on phenolics in woody plants (2) costs of plant antiherbivore defenses	Nonlinear approx. 3-fold decrease in magnitude of effect sizes with publication year in both datasets
Nykänen and Koricheva 2004	Damage-induced changes in woody plants	Dramatic nonlinear decrease in effects of damage on phenolic and nutrient concentrations in host plants and on herbivore performance with publication year
Møller et al. 2005	Relationship between fluctuating asymmetry and sexual selection	No difference in mean effect size between studies conducted through 1996, studies conducted in 1997–2001, and the sample of unpublished studies in the present study
Saikkonen et al. 2006	Endophyte-grass interactions	Temporal decrease in effects of endophytes on plant competition, plant performance, and resistance to herbivores
Toth and Pavia 2007	Induced herbivore resistance in seaweeds	Large decrease in magnitude of induced resistance from the late 1980s to early 2000s
Zvereva et al. 2008	Effects of air pollution on species richness of vascular plants	Effect size did not change with the publication year

(*continued*)

TABLE 15. 1. *Continued*

Reference	Topic	Pattern revealed
Kampichler and Bruckner 2009	Role of microarthropods in litter decomposition	Significant decrease in the magnitude of the effect size with the year of publication
Barto and Rillig 2010	Effects of plant herbivory on mycorrhizae	Significant decrease in the magnitude of the effect size with the year of publication
Krist 2010	Egg size and offspring quality in birds	Significant decrease in the magnitude of the effect size with the year of publication
Santos et al. 2011	Dominance and plumage traits	Significant decrease in the magnitude of the effect size with the year of publication
Kelly and Jennions 2011	Sexual selection and sperm quantity	Effect size did not change with the publication year

hypothesis-driven research, and the majority of studies reported a decrease rather than an increase in the magnitude of the effect size with publication year (but see Alatalo et al. 1997). Reported changes in the magnitude of the effect sizes are often quite dramatic (e.g., 2- to 3-fold or more), and sometimes lead to the loss of the statistical significance of the mean effect size, or even to a change in the sign of the effect size (Leimu and Koricheva 2004, Nykänen and Koricheva 2004, Saikkonen et al. 2006). Initially, these temporal trends were treated as isolated occurrences and attributed to paradigm shifts (sensu Kuhn 1970), scientific fads, changes in methodological approaches, or biases in the choice of study systems. However, Jennions and Møller (2002b) have analyzed 44 independent meta-analytic data sets covering a wide range of ecological and evolutionary topics, and found a small but significant decrease in effect size with year of publication across these data sets. It is clear, therefore, that temporal trends in the magnitude of effect sizes represent a general phenomenon in ecology and evolutionary biology, and the most common pattern appears to be a decrease in effect sizes with time.

EVIDENCE OF TEMPORAL TRENDS IN EFFECT SIZES IN OTHER RESEARCH FIELDS

Temporal trends in effect sizes have also been repeatedly reported in clinical medicine. For example, Trikalinos et al. (2004) found that the magnitude of the effect size of therapeutic and preventive interventions in mental health has changed considerably over time; similarly to ecology and evolutionary biology, for three out of four response variables of outcome, a decrease in effect size was more common than an increase. Furthermore, in eight out of 100 meta-analytic data sets examined, the statistical significance of the mean effect size was lost as more trials were published. Similarly, Gehr et al. (2006) demonstrated fading of reported effectiveness over time in three out of four investigated lipid-lowering and anti-glaucoma drugs.

Ioannidis (2005a) showed that the results of 32% of highly cited original clinical research studies that he examined were either contradicted by subsequent research or found stronger effects than subsequent studies. In another study, provocatively titled "Why most published research findings are false," Ioannidis (2005b) argued that the probability of a research finding being false (and thus the probability of this finding being refuted by subsequent research) is particularly high if the study's sample size is small, the effect size in a research field is small, and the scientific field is hot.

In molecular genetics research on genetic associations, a rapid early sequence of extreme, opposite results was observed (Ioannidis and Trikalinos 2005). This phenomenon was named the "Proteus phenomenon" after the mythological god who rapidly metamorphosed himself into different figures. Furthermore, in 21 out of 36 studied meta-analyses of associations in genetic epidemiology, the first study or studies tended to give more impressive results (Ioannidis et al. 2001). Ioannidis and Trikalinos (2005) suggest that the Proteus phenomenon might be characteristic of disciplines where data production is rapid and copious (like "omics" fields), but might be less likely in research where studies take considerable time to perform, as is the case in many ecological field studies and clinical trials in medicine.

Temporal trends in effect sizes have also been observed in the social sciences. For example, the reported magnitude of gender difference in cognitive abilities (Feingold 1988), mathematics performance (Hyde et al. 1990), and sexual attitudes and behaviors (Olivier and Hyde 1993) has declined over time, whereas the effects of the media on body image concerns among American women have increased in the 2000s relative to the 1990s (Grabe et al. 2008).

Recently, Ioannidis (2008) reviewed theoretical work and empirical evidence of decreases in effect sizes over time and suggested that this is a general phenomenon across scientific fields, and that effect sizes in early studies on the topic are often inherently inflated. But why is this the case?

POSSIBLE CAUSES

Early studies are prone to overestimate the magnitude of the effect if they rely on statistical significance testing to establish the existence of an effect, and are based on small sample sizes (Ioannidis 2008). For example, Schmidt (1992) showed that if the actual effect size (standardized mean difference) is equal to 0.5, for a study with a sample size $n = 30$ (equal to statistical power of 0.37) to be significant at $P = 0.05$, the reported effect size must be 0.62 or larger, which is 24% larger than the real effect size (0.50). Furthermore, the average of the significant effect size values in the above example would be 0.89, which is 78% larger than the true value. The lower the sample size and, thus, the lower the statistical power of the early studies, the more likely these are to overestimate the magnitude of the effect (Ioannidis 2008). Ioannidis and Lau (2001) have examined changes in treatment effects over time in two medical fields (pregnancy/perinatal medicine and myocardial infarction). They have shown that the probability that the effect size changes with the accumulation of more data is a function of the cumulative number of patients. It might be expected that stabilization of effect sizes around the mean across studies will take even longer in ecology since unlike medicine where the study subject is a single species (*Homo sapiens*), the diversity of study subjects in biology is much larger and sample sizes are typically small. However, Koricheva et al. (in preparation) have found no evidence of changes in sample sizes over time across 54 meta-analytic data sets on various topics in ecology and evolution. Therefore, earlier studies in ecology and evolutionary biology tend to have the same degree of replication as later studies on the same topic, and thus temporal trends in effect sizes in these fields cannot be explained by lower statistical power of earlier studies.

Jennions and Møller (2002b) suggested that the most general and plausible cause of the observed decrease of effect sizes with time in ecology and evolutionary biology is time-lag bias, the delayed publication of studies reporting small or statistically nonsignificant effects. When time-lag bias operates, the first published studies will report larger effect sizes compared to subsequently published investigations, resulting in a decrease in the mean effect size over time. Such bias has been shown to occur in clinical medicine (Stern and Simes 1997, Ioannidis 1998) as well as in genetic association studies (Ioannidis et al. 2001), but at present there is no direct

evidence of time-lag bias against nonsignificant results in ecology (Chapter 14). In contrast, Koricheva (2003) has found that ecological studies with a large proportion of nonsignificant results are more likely to be published than studies with a smaller proportion of nonsignificant results because the latter are submitted to journals with larger impact factors, which have higher rejection rates. She also found no evidence of delayed publication of studies with a large proportion of nonsignificant results.

Another form of selective reporting which might contribute to temporal trends in effect sizes is deliberate withholding or delayed publication of studies that fail to confirm the hypothesis being tested. This type of bias is especially likely to occur in early studies testing a recently suggested hypothesis because of the initial enthusiasm and less critical attitude of scientists toward new and currently popular ideas (Kuhn 1970). Gradually, however, evidence refuting the hypothesis begins to accumulate and, eventually, alternative scenarios and competing theories are suggested; this prompts publication of studies supporting these new theories and leads to temporal changes in magnitude or even sign of the overall effect (Leimu and Koricheva 2004). Evidence that this type of bias is responsible for some of the temporal trends in effect sizes reported in ecology comes from several sources. For example, an increase in reported heritability estimates of secondary sexual characters began after new models that appeared between 1986 and 1988 indicated that such characters can be honest indicators of male viability, providing fitness benefits for choosy females (Alatalo et al. 1997). Similarly, a decrease in the magnitude of reported fitness costs of plant resistance to herbivores began in the early 1990s, after several studies providing theoretical justification for the absence of fitness costs were published (Leimu and Koricheva 2004). The above observations have led to the suggestions that "analyses aimed at assessing the generality of recently advanced paradigms should wait until revolutions have settled" (Simmons et al. 1999). Further evidence for the role that bias against nonconfirmatory evidence plays in temporal changes in effect comes from the study by Poulin (2000). Poulin found that a temporal decrease in effects of parasites on host behavior occurred only among studies that were specific tests of the adaptive manipulation hypothesis, but was not apparent among more descriptive studies that examined the effects of parasites on host behavior in terms of pathology or other consequences for the host. In some fields, however, highly contradictory findings might be more attractive to investigators and editors, resulting in a rapid succession of extreme opposite results; this has been observed in molecular genetics (Ioannidis and Trikalinos 2005).

Temporal changes in effect sizes might also be caused by a bias in the choice of study organisms or systems. Gurevitch and Hedges (1999) suggested that ecologists tend to perform their studies on organisms that are more likely to display statistically significant responses; they called this tendency a "research bias." As the range of study organisms tested increases with time, the magnitude of the cumulative effect sizes diminishes. For example, Tregenza and Wedell (1997) suggested that the increase in published heritability estimates of secondary sexual characters observed by Alatalo et al. (1997) could be due to an increase in the number of studies conducted on birds, in contrast to earlier studies that were conducted largely on insects. Nykänen and Koricheva (2004) have demonstrated that the decrease in magnitude of the reported effects of plant damage on herbivore performance (indicating induced resistance) was partly due to a decrease in the proportion of studies conducted on mountain birch, which manifested stronger induced resistance than other tree species. Similarly, Saikkonen et al. (2006) have shown that most of the conceptual framework for endophyte-plant interactions has been based upon studies of two economically important grass species (tall fescue and perennial ryegrass), particularly tall fescue cultivar Kentucky 31. This cultivar is, however, a misleading model system for endophyte-grass interactions because it performs better than other cultivars

and plants collected from nature. As the diversity of study systems increases over the years, the cumulative effect of endophytes on plant competition, performance, and resistance to herbivores decreased; it eventually became nonsignificant in the case of effects on plant competition and performance.

Another potential cause of temporal changes in effect sizes are changes in research or statistical methods over time. For example, Simmons et al. (1999) suggested that the temporal decline in the proportion of studies supporting the role of fluctuating asymmetry in sexual selection was due to an increase in the proportion of studies that used repeatability analysis to distinguish fluctuating asymmetry from measurement error. Changes in statistical analyses, such as better control for confounding variables, could potentially have the same effect. Similarly, as another explanation for the observed temporal increase in heritability estimates of secondary sexual traits observed by Alatalo et al. (1997), Tregenza and Wedell (1997) pointed out that before 1988 the majority of studies used artificial selection; however, after 1998 there was a marked increase in studies using parent-offspring or sib-sib regression in the wild. More recently, Timi and Poulin (2007) demonstrated how changes in the analytical methods used to study patterns of species composition in parasite communities resulted in increases in the likelihood of finding nestedness over time. Such changes in research or statistical methods are common in ecology as well in many scientific fields. This could potentially account for temporal changes in effect sizes observed in some studies, and might results in either increases (Timi and Poulin 2007, Barto and Rillig 2010) or decreases (Simmons et al. 1999) in effect sizes. (See also Worked example 1.)

Since the majority of effect size metrics represent comparisons of frequencies or means between a control and treatment groups (e.g., odds ratios, response ratios, standardized mean differences), temporal changes in event rate or magnitude of the mean in the control group could also account for temporal changes in effect sizes. In medicine, effects of the control rate (the proportion of patients in the control group with the event of interest) on treatment efficacy are well known (Schmid et al. 1998). Temporal changes in the control rate could be due to many factors, such as improved standards of medical care and diagnostic tools, public awareness of the need to get care sooner, and so forth. For example, Antman and Berlin (1992) reported a decrease in the incidence of ventricular fibrillation (VF, cardiac muscle arrhythmia) in patients with acute myocardial infarction (AMI), presumably as a result of dramatic improvements in the general care of AMI patients since the 1960s; the authors pointed out the need to reassess the risk-benefit ratio of using lidocaine prophylaxis treatment for VF, especially in view of a previously reported trend toward excess mortality in lidocaine-treated patients. Similarly, Gehr et al. (2006) also demonstrated that a fading reported effectiveness of several pharmaceuticals could be explained to a large extent by the decrease in the baseline values of the parameter of interest (i.e., patients who had been included in the earlier trials were sicker than patients in later trials).

In ecology, temporal changes in control rates and resulting changes in effect sizes might be expected. Examples include studies assessing losses of biodiversity, habitat fragmentation, and global climate change; as the causes of these responses continue to change in frequency or extent, we might see corresponding changes in the magnitude of the effect sizes assessing these responses over time. Gardner et al. (2003) showed that the rate of coral loss (as measured by the annual rate of change in percent coral cover) in the Caribbean basin decreased in most areas during the 1990s, compared to the 1980s. This decrease in the effect size might suggest alleviation of some of the pressures causing coral mortality. However, it could also indicate that the remaining types of corals are hardier and less sensitive to human-caused disturbance than the corals that disappeared first. More pessimistically, the decrease in rate of coral loss could

be a "control rate" type effect, and simply reflect the fact that there is relatively little coral left to lose and thus, as coral cover approaches zero, the rate of coral loss is expected to slow down.

Finally, in some cases temporal changes in effect size might reflect real biological phenomena and be due to rapid adaptation to changes in the strength or direction of the selection pressure. A well known example in medicine is the development of resistance to drugs by bacteria and viruses, which might decrease treatment efficacy for the same treatment over time (Fischbach et al. 2002). Similar adaptive responses may occur in ecological and evolutionary studies; examples include the response to selection pressure imposed by herbicide and pesticide application, or overharvesting by humans (Strauss et al. 2008). Furthermore, Strauss et al. (2008) suggest that the evolutionary history of the study population prior to the experimental manipulation may also strongly influence both the initial magnitude of the treatment effect and the trajectory of subsequent evolution.

To summarize, various factors might explain temporal changes in effect sizes across studies, and several of them might be operating in each particular case. Temporal changes in effect sizes are often indicative of other sources of heterogeneity that might have been missed initially if temporal trends were not examined; exploration of temporal trends in effect sizes is thus a useful diagnostic tool to reveal those causes of heterogeneity. In order to demonstrate that the observed temporal changes reflect real changes in the magnitude of the biological effect over time, the researchers have to rule out other possible explanations, such as publication bias (Chapter 14), heterogeneity between study organisms (Chapter 17), and changes in research methods (see Worked example 1).

METHODS OF DETECTION

Graphical methods

The simplest way to visualize a potential temporal trend in a meta-analytic data set is to produce a scatterplot of effect sizes versus publication year (Figs. 15.1A and 15.2A). Another graphical technique, called cumulative meta-analysis (CMA), was introduced to examine temporal trends in effect sizes in medicine (Lau et al. 1992). In order to conduct CMA, one has to sort individual studies in chronological order, and the earliest available study is then entered into the analysis first. At each step of the CMA, one more study is added to the analysis and the new mean effect size and 95% confidence interval are recalculated. This allows estimation of the contribution of individual studies and assessment of temporal change in the magnitude and direction of research findings. In the absence of biases and heterogeneity, the CMA plot should exhibit a fairly constant estimate of treatment effects over time, with some fluctuations due to chance only in the early steps. As more studies are added to the analysis, the cumulative effect size stabilizes around the mean, and the width of the confidence intervals decreases.

Originally, CMA was proposed as a tool in medicine to detect the earliest year at which a treatment effect became statistically significant, and thus a conclusion about its clinical efficiency could be drawn and a decision about its use made. For example, by using CMA, Lau et al. (1992) demonstrated that the evidence in favor of streptokinase drug therapy for patients with myocardial infarction became significant 13 years before the experts recommended its widespread use. Thus, patients continued to receive inferior treatment long after the evidence was available to demonstrate that other treatments were more effective. More recently, Fergusson et al. (2005) reviewed the results of 64 clinical trials of aprotinin, a serine protease inhibitor used to reduce bleeding during cardiac surgery; the trials were conducted between 1987 and 2002. They showed that the use of CMA would have allowed establishing a clinically significant effect of the drug after 12th trial published in 1992, making the subsequent

42 trials redundant, unethical in regard to the patients, and wasteful of time and resources. Interestingly, Fergusson et al. (2005) revealed that a large number of redundant trials evaluating efficacy of aprotinin were conducted because researchers were not adequately citing previous research.

Mullen et al. (2001) suggested that CMA could be used to assess sufficiency and stability of cumulative knowledge. Consideration of sufficiency addresses the question: "Are additional studies needed to establish the existence of the phenomenon?" If the answer to the above question is no, then collecting additional evidence for an already established effect might waste time and resources, and delay implementation of effective treatments (in medicine) or conservation management policies (in conservation biology). The consideration of stability addresses the question: "Will additional studies change the evidence of the phenomenon's existence and strength?" This aspect directly relates to temporal changes in effect sizes; if the cumulative mean effect on the CMA plot keeps changing with each new study added to the analysis, the results of the meta-analysis should be interpreted with caution because new evidence might change those conclusions. Moreover, instability of the CMA plot might suggest that an important source of heterogeneity exists among the studies, and thus calculation of the mean effect across the whole data set is less meaningful.

CMA is usually applied to a collection of studies retrospectively, to check whether and when the evidence for phenomena under consideration has achieved sufficiency and stability (e.g., Lau et al. 1992). However, Mullen et al. (2001) also recommended CMA as a prospective tool for newly emerging topics with relatively few studies available. They argued that such a prospective approach would help inform researchers about the necessity for investing additional resources in conducting studies on the topic when sufficiency and stability remain uncertain. Mullen et al. pointed out that this prospective approach to CMA removes a commonly raised objection to meta-analysis, namely that the compiled database is too small for a meta-analysis; the prospective approach therefore makes the application of meta-analysis to new research fields imperative rather than suspect.

While the application of CMA in medicine is now widespread, this method has been introduced into ecology only recently (Leimu and Koricheva 2004), even though it is available in MetaWin (Rosenberg et al. 2000), the meta-analysis software package most widely used by ecologists. Somewhat alarmingly, the first applications of CMA in ecology detected severalfold changes in the magnitude of the effect (Leimu and Koricheva 2004, Toth and Pavia 2007), losses of statistical significance of the effects reported in earlier studies (Saikkonen et al. 2006), and even changes in the sign of the cumulative effect size over time (Nykänen and Koricheva 2004). Future applications of CMA to a larger number of ecological data sets will reveal whether these are extreme cases or standard patterns as in studies of genetic associations, where every possible temporal pattern for cumulative effect trajectories has been observed (Ioannidis et al. 2001).

Note that in CMA, data can be arranged not only in chronological order but also in order of any other continuous variable or covariate of interest to the reviewer— for example, by study sample size or control rate (Lau et al. 1995), or by impact factor of the journal in which the study is published (Leimu and Koricheva 2004). In addition, CMA can be conducted on subgroups of studies to take into account heterogeneity in study organisms or in research methods so that one can compare temporal patterns across these subgroups. This is important, because heterogeneity in the data might cause spurious temporal patterns in effect sizes (see Worked example 1).

Several meta-analytic statistical software packages, such as MetaWin and Comprehensive Meta-Analysis, include an option to conduct CMA and to produce a CMA plot (see Chapter 12

and worked examples). However, note that in MetaWin, if a random-effects model is chosen for CMA, it will automatically switch to calculate a fixed-effects model if the between-study variance estimate is less than or equal to zero. Because the estimates of between-study variance might vary at each step of CMA when a new study is added, the resulting CMA plot in MetaWin will often represent a mixture of mean effects and 95% confidence intervals calculated on the basis of fixed- and random-effects models. Because random-effects models usually produce broader confidence intervals, it might be difficult to draw conclusions about the convergence of effect size from such plots. CMA plots produced by the Comprehensive Meta-Analysis software are free from this problem.

Ioannidis and Lau (2001) suggested an extension of CMA, a recursive cumulative meta-analysis (RCMA), which shows the relative change in the magnitude of the treatment effect as a new study is added to the meta-analysis. The relative change at each step of the cumulative analysis is calculated as E_{t+1}/E_t, where E_t and E_{t+1} are the cumulative mean effect sizes at steps t and $t+1$, respectively. The benefit of RCMA is that results from several cumulative meta-analyses using different metrics of effect sizes and reporting different magnitudes of effects can be plotted on the same graph to compare the patterns (Ioannidis and Lau 2001). Observed relative changes in the magnitude of the cumulative effect size reflect the uncertainty of the treatment effect. Moreover, Ioannidis and Lau (2001) showed that early fluctuations in the magnitude of the treatment effect might sometimes signal further major changes in the magnitude of the effect sizes. Note, however, that recursive meta-analysis based on relative change cannot be used in situations where the sign of the effect size varies between different information steps (t and $t+1$) of CMA, which is often the case for Fisher's z, Hedges' d, and the log response ratio. In these cases, one can use the absolute difference ($E_{t+1} - E_t$) instead. Furthermore, RCMA works best for symmetric effect sizes such as odds ratios, where the result of the analysis would be the same if E_t/E_{t+1} were used instead of E_{t+1}/E_t (Trikalinos and Ioannidis 2005); this is not true for nonsymmetric effect sizes, such as the relative risk. In view of the above limitations and the prevalence of metrics other than odds ratios in ecology and evolutionary biology, RCMA may prove to be of limited use in these fields.

Statistical methods

Graphical tools like scatterplots or CMA plots are useful for initial inspections of data but, as all visual methods, they might be subject to misinterpretation (compare with funnel plots, see Chapter 14) and should be supplemented by formal statistical methods. In addition, from a frequentist perspective CMA suffers from the problem of multiple testing of the same hypothesis and an inflated type I error (Bender et al. 2008). An infinitely updated CMA would eventually yield a statistically significant finding even when the true effect size is 0 (Berkey, Mosteller et al. 1996). Several techniques have been proposed to adjust P-values and test statistics to multiple testing in CMA (Pogue and Yusuf 1997, Lan et al. 2003). Some authors argue, however, that accumulating meta-analyses are best interpreted in a Bayesian framework and that there is no need to adjust for multiple testing in CMA (Lau et al. 1995). Bender et al. (2008) suggest that the relevance of adjusting for multiple testing in CMA depends on whether the review is intended for descriptive or decision-making purposes. If the former, the adjustment is not required, but if the latter, an adjustment might be advisable. The recently proposed sequential approaches (Brok et al. 2008, Wetterslev et al. 2008, Higgins et al. 2011) may reduce the risk of false positives in cumulative meta-analysis by using the approach analogous to sequential monitoring boundaries. This method deserves attention in future research and we refer interested readers to the above publications, which describe the method in more detail.

Various statistical tests can be conducted to assess the significance of the temporal changes. For example, one can compare whether the results of the first study (studies) differ more than expected by chance from those of the subsequent studies on the topic by using the formula:

$$z = \frac{T_1 - \overline{T}_2}{\sqrt{v_1 + v_2}}, \tag{15.1}$$

where T_1 is the effect size of the first study (studies), T_2 is the mean effect size of all subsequent studies, and v_1 and v_2 are their corresponding variances (Trikalinos and Ioannidis 2005). Absolute values of $z > 1.96$ indicate statistically significant differences at the 5% significance level. The above formula works for effect sizes that follow an approximate normal distribution. It should therefore be used on Fisher's z rather than Pearson's correlation coefficients, and on log response ratios and log odds ratios rather than response ratios and odds ratios, respectively.

One can also use a nonparametric test (e.g., the sign test) to compare the number of steps in cumulative meta-analysis where the effect size increases rather than decreases (Trikalinos and Ioannidis 2005). The assumption is that in the absence of bias, the number of steps in which effect size is increasing and decreasing should be equal. However, in our experience such tests are very conservative because they take into account only the direction, but not the magnitude, of change in the effect size at each step. Even if each step where a decrease in effect size occurs results in a much larger change in effect size than each step where an increase occurs (resulting in a visible decrease in effect size on the CMA or regression plot), no significant difference will be detected by the test; however, this is only the case if the number of steps where decreases and increases occur are similar. Therefore, we do not recommend the use of this test.

Mullen et al. (2001) have suggested quantitative indicators of sufficiency and stability that can be derived from CMA. As an indicator of sufficiency, they recommend calculation of the fail-safe ratio, which simply indicates whether the fail-safe number (Chapter 14) for the current step of CMA exceeds the benchmark of $5N + 10$. If and when the fail-safe ratio exceeds 1, it indicates that cumulative weight of evidence is sufficiently tolerant for future null results. This approach suffers from the same shortcomings as calculation of a fail-safe number (Chapter 14). Some of these problems can be alleviated by using the weighted method of fail-safe sample size (N_{fs}) calculation (Rosenberg 2005); however, the N_{fs} method still severely overestimates the number of studies needed to make the magnitude of the effect nonsignificant if missing studies report an effect with an opposite sign rather than null results, as is often the case in ecology. In addition, calculation of N_{fs} makes sense only if the magnitude of the effect size is significantly different from zero, and thus it cannot be applied to steps in CMA where the effect size is not significant. This limits its usability as sufficiency can never be reached in meta-analysis where the mean effect size is not significant.

As an indicator of stability, Mullen et al. (2001) suggest calculation of the "cumulative slope," which is the slope of the regression of cumulative effect sizes from all the previous and current steps of CMA, along with each new step in CMA. Stability is achieved when the slope of the regression approaches 0, indicating that adding another study causes little change in the cumulative effect size. Note that the significance of the slope cannot be formally estimated because of the problem of multiple testing and the fact that meta-analysis data violate the assumptions of the general linear model for statistical inference. Therefore, even though the estimate of the slope itself is not biased, the decision as to when the slope becomes negligibly small remains somewhat arbitrary.

One can also conduct linear weighted regression analysis for the relationship between effect sizes and publication year (Chapters 8 and 9; Gehr et al. 2006). This method captures temporal trends well when the magnitude of the effect size exhibits a uniform and monotonous decrease

or increase with time. However, this is not always the case, and uneven, irregular shifts in effect size in opposite directions have been observed both in ecology (e.g., Nykänen and Koricheva 2004) and in molecular genetics research on genetic associations (Ioannidis and Trikalinos 2005). Two alternative curve-fitting methods, fractional polynomial regression and spline regression, have been suggested to quantify nonlinear associations in meta-analysis (Bagnardi et al. 2004).

Finally, Kulinskaya and Koricheva (2010) have recently proposed the use of statistical quality control (QC) charts, in particular CUSUM charts (Hawkins and Olwell 1997), to assess significance of effects and detect trends over time in cumulative meta-analysis. Methods of statistical quality control were initially developed in industrial applications of statistics to assess whether the variability of a production process was due to chance or to assignable causes. When there is no temporal shift, the process is in control and all effect estimates are normally distributed with the same mean. If a shift happens at some point in time, the mean of the process deviates from the mean, and the process can be considered out of control. Nowadays, quality control charts are commonly used in medicine, epidemiology, and public health to detect a start of an epidemic or to control quality within the United Kingdom's National Health Service. The QC procedures are available in most major statistical packages, including R. Kulinskaya and Koricheva (2010) illustrate the use of QC charts for detection of temporal trends by using several examples, including the meta-analysis by Torres-Vila and Jennions (2005), which is used as one of the data set examples in this book.

If evidence of temporal changes in effect sizes is obtained, the next step is to examine underlying causes of this trend. Barto and Rillig (2010) used a simple approach to find out which moderators changed with publication year; this could then explain the observed temporal patterns in studies of herbivory effects on mycorrhizae. For each potential moderator (e.g., research method used or taxonomic group studied), they have sorted the studies in chronological order and established which level of moderator was used in the earliest study. This level was assigned the value of 1, the next level of moderator was assigned the value of 2, and so on. The authors then performed correlation analyses between the levels of each moderator and the year of publication; this analysis revealed that temporal changes in effect size were most likely to be due to changes in the type of plants used in experiments and the treatment methods.

WORKED EXAMPLES

1. Effects of elevated CO_2 on net CO_2 assimilation in woody plants

As the first example, we have selected a subset of studies from the database compiled by Peter Curtis (Curtis 1996, Curtis and Wang 1998), and reporting the effects of elevated CO_2 on net CO_2 assimilation in woody plants (response variable PN) from 39 studies published over a 10-year period (1987–1996). The effect size metric used in this meta-analysis is the log response ratio. Positive values of the effect size indicate an increase in net CO_2 assimilation in plants under elevated CO_2 as compared to ambient CO_2 levels. In the original database, several individual studies contribute more than one data point because they report results for different plant species. In order to begin the analysis of temporal trends, we have to first average effect sizes and their variances by study because results within the same study, even if considered relatively independent as in the case of different plant species, have been published simultaneously; it does not make sense to add them sequentially in cumulative meta-analysis.

Once we have calculated mean effect sizes and variances per study, we can subject the data to a formal meta-analysis. The overall mean effect size is 0.369 and it is significantly different from 0, as indicated by the 95% confidence intervals. The heterogeneity analysis using the

fixed-effects model yields a total heterogeneity estimate of $Q_t = 1018$, which is much larger than 70.703, the critical chi-square value for $df = 38$ ($P = 0.001$). This suggests that between-study variation is significantly larger than would be expected from sampling error alone. We can therefore proceed to search for moderators.

Even though the studies included in this meta-analysis have been conducted over a relatively short time period (10 years), important methodological developments in terms of exposure facilities have taken place during this time (Curtis 1996, Curtis and Wang 1998). It might be interesting to examine, therefore, whether any temporal trends in effect size are apparent. In order to examine whether the magnitude of the effect size changes with time, we first produced a scatterplot of effect size against publication year (Fig. 15.1A). The plot appears to indicate an increase in the magnitude of the effect with publication year. Figure 15.1B shows the CMA plot for the same data set based on the random-effects model. Visual inspection of the CMA plot reveals that the cumulative effect size is increasing with time from being weak (and not significantly different from 0) in early studies, to being significantly positive by the end of the meta-analysis. This indicates an increase in CO_2 assimilation in response to elevated CO_2 concentrations. Overall, the magnitude of the effect size changed more than 2-fold over 10 years, but the increase in effect size was nonlinear and most of the changes took place during the first 10 experiments. However, the magnitude of the effect changed little in the last 20 studies, although the confidence interval kept getting smaller.

In this data set, more than one study on the topic was published each year, except in 1988 and 1996 (Fig. 15.1A). It is difficult to determine the exact chronological order in which such studies were conducted and published (especially given the practice of some journals to provide early online access to articles which will be included in future issues). Yet, the order in which studies published in the same year are entered in the cumulative meta-analysis might affect the shape of the CMA plot (on the CMA plot in Fig 15.1B the order is alphabetic). Therefore,

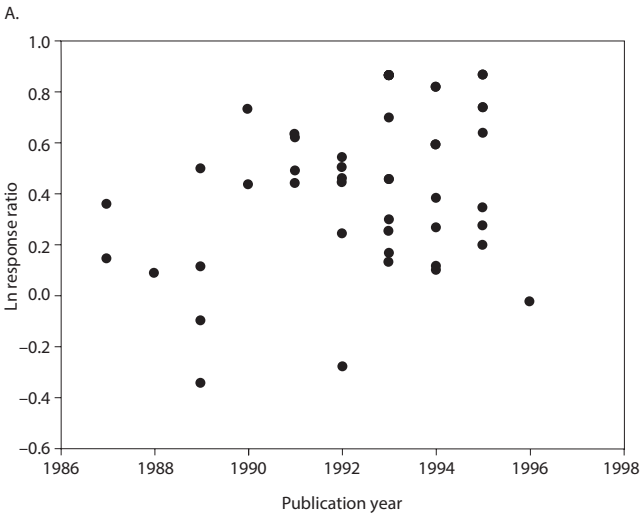
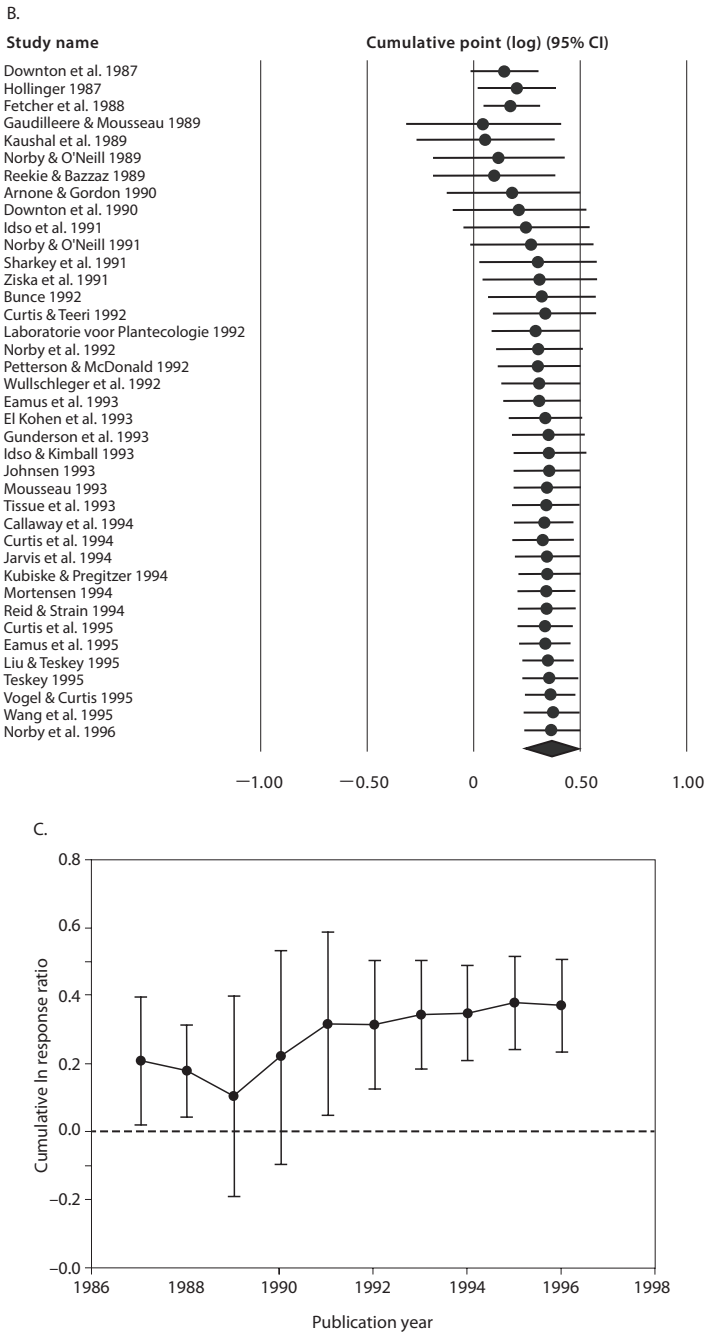


Figure 15.1. Scatterplot (*A*), cumulative meta-analysis plot (*B*), and cumulative meta-analysis trajectory (*C*) of studies examining effects of elevated CO_2 on net CO_2 assimilation in woody plants (data from Curtis and Wang 1998). The CMA plot was plotted using Comprehensive Meta-Analysis software.



one might prefer to draw the CMA plot based not on study-specific effect sizes, but on year of publication. We can do this by using the values of cumulative means and 95% CI at the end of each calendar year regardless of how many studies have been published in that year. This gives the trajectory of change in cumulative effect size and confidence intervals over the years (Fig. 15.1C). The CMA trajectory plot shows that there is a clear difference in magnitude of effect

sizes before and after 1991, thus supporting the conclusion that a nonlinear temporal change in effect sizes took place.

We then performed a weighted least square regression with publication year as an explanatory variable. Since between-study variation is significant, as indicated by heterogeneity analysis, using the random-effects model is more appropriate. This analysis can be conducted by using either the Comprehensive Meta-Analysis software (by selecting method of moments computational option) or the MetaWin software (by selecting continuous random model option in the summary analysis menu). A separate column containing publication year has to be created in the data file and selected as a moderator/predictor. Both of the software packages produced identical results. Variation in effect sizes explained by the model was not significant ($Q_M = 1.86$, $df = 1$, $P = 0.172$), neither was the slope (0.031) nor the intercept (-63), suggesting that publication year is a poor predictor of elevated CO_2 effects on net CO_2 assimilation in woody plants. This conclusion appears to disagree with the results of the CMA discussed above. Recall, however, that the linear regression model that we have applied assumes a linear relationship between the dependent and independent variable, and both the scatterplots and CMA plots suggest that the relationship between publication year and effect size is nonlinear. Therefore, the use of alternative regression methods which allow quantifying non-linear associations in meta-analysis (e.g., Bagnardi et al. 2004) would be preferable in this case.

We could also estimate whether the effect size of the first study in this analysis differed more than would be expected by chance from the results of all subsequent research. The effect size of the first study (Downtown et al. 1987) is 0.148 and variance is 0.007 (Fig. 15.1A). Mean effect size of all other studies excluding Downton et al. could be easily calculated with the Comprehensive Meta-Analysis software by selecting the "one study removed" option (or by removing the row containing the results of the first study from the data file and repeating the meta-analysis). Based on random-effects models, it gives an effect size of 0.375 and a variance of 0.005. By using Equation 15.1, $z = \frac{0.148 - 0.375}{\sqrt{0.007 + 0.005}} = -2.072$, which is larger than the critical value of 1.96 for $P = 0.05$, suggesting that the results of the first study by Downton et al. (1987) differed beyond chance from the results of the subsequent studies.

What could be the cause of the observed increase in the magnitude of the effect with publication year? It cannot be explained by publication bias against nonsignificant results, and publication bias against studies reporting significant negative effects of elevated CO_2 on net assimilation appears unlikely. Neither the ambient and elevated levels of CO_2, nor duration of the exposure, changed significantly over the examined period of 1987–1996. However, previous meta-analyses of the same database by Curtis (1996) and Curtis and Wang (1998) revealed that exposure facility and pot size (rooting volume) significantly affect plant responses to elevated CO_2 in terms of net CO_2 assimilation. Studies conducted indoor in controlled-environment growth chambers (GC) reported smaller effects on net assimilation than studies conducted in greenhouses (GH) and in the field-based open top chambers (OTC), presumably because of the lower light levels in GCs. The use of GCs decreased with time from 35% of all studies conducted between 1987 and 1991, to 21% of studies conducted in 1992–1996, and the use of GH/OTC facilities increased from 65% to 79% for the same time periods. In addition, Curtis and Wang (1998) have shown that net CO_2 assimilation was significantly lower in experiments where plants have been grown in smaller pots (< 2.5 L) than in plants grown in larger pots or in the ground. Only 1 experiment published in 1987–1991 was conducted in the ground as compared to 14 experiments in the ground in 1992–1996. Among the experiments conducted in the pots, average pot size increased from 3 liters in 1987–1991 to 6 liters in 1992–1996. Therefore, methodological changes in exposure facilities and pot size are likely to contribute to the observed temporal changes in net CO_2 assimilation. Although the above

sources of heterogeneity have been revealed in previous meta-analyses by Curtis (1996) and Curtis and Wang (1998), it has not been shown before that these methodological changes may lead to temporal trends in effect sizes. Note that if one conducted a meta-analysis on effects of elevated CO_2 on net CO_2 assimilation in 1990, after the first 9 studies on the topic were published, the magnitude of the mean effect size obtained (0.219) would be only 60% of that obtained by 1996 (0.369). Therefore, in the presence of temporal changes in effect sizes, early meta-analyses might considerably over- or underestimate the magnitude of the effect size.

2. Effects of male mating history on female reproductive output in Lepidoptera

The second example is based on a meta-analysis by Torres-Villa and Jennions (2005) that compared reproductive output of female Lepidoptera that mated with virgin males, with those that mated with experienced males. The data set includes 29 studies published in 1971–2003. The metric of effect size used is Hedges' *d* and positive effects indicate higher reproductive potential of females that mated with virgin males. In this data set, each study contributes only one effect size to the analysis, so there is no need to average the effect by study. The overall mean effect size for 29 studies is $d = 0.335$ and is significantly different from 0, suggesting that females had higher reproductive output when mated with virgin rather than with experienced males. The heterogeneity analysis using a fixed-effects model yields a total heterogeneity estimate of $Q_t = 57.48$, which is significant at $P = 0.001$, suggesting that between-study variation is significantly larger than would be expected from sampling error alone; we thus can proceed to search for moderators.

We have no a priori reasons to suspect that the magnitude of reproductive benefits derived by females mating with virgin males is different in studies conducted in the 1970s than those reported in later studies. However, because the time span of the studies included in this analysis is quite extensive (over 30 years), it might be prudent to examine whether the magnitude of the effect size changes with time. We first produced a scatterplot of effect size against publication year (Fig. 15.2A), which revealed no obvious temporal trend apart from the very last study reporting a larger estimate than all the previous ones. The CMA plot and trajectory (Fig. 15.2B–C) do not reveal considerable temporal variation in the effect size either. There is fluctuation in the cumulative effect size from the first 7 studies, but then the cumulative effect quickly converges to the mean value and remains stable in the last 20 studies.

Note that CMA is much less sensitive to changes in effect sizes that occur in later studies. For example, the last study in this meta-analysis reported an effect size which is two times larger than any effect reported in previous studies (Fig. 15.2A), but the cumulative mean effect barely changes (Fig. 15.2B–C). This is because to change the pooled effect size at later stages of CMA, new studies have to overcome an increasing amount of inertia (accumulating evidence). This is not a big problem if the majority of temporal changes take place early (as is often the case, compare with Worked example 1). However, if the scatterplot suggests that temporal changes take place mainly at the most recent time interval, one might want to conduct CMA in reverse chronological order (i.e., entering the most recent studies first).

To statistically test for the presence of a temporal trend, we have also performed a weighted least square regression, with publication year as an explanatory variable. A random-effects model was used; variation in effect sizes explained by the model was not significant ($Q_M = 0.14$, $df = 1$, $P = 0.709$), nor was the slope (-0.003) or the intercept (5.49). The *z* test shows that the results of the first study do not differ from the subsequent research more than would be

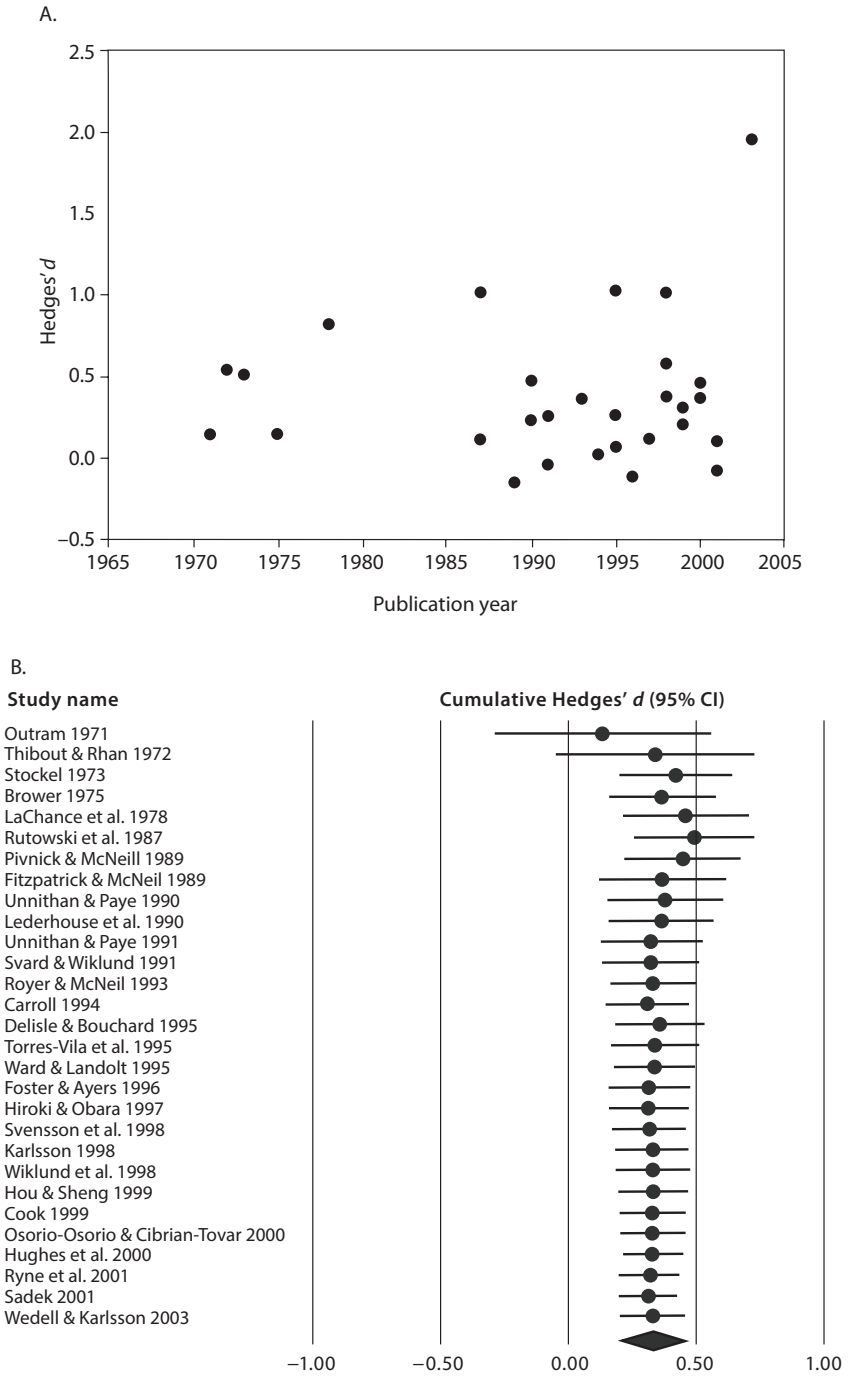
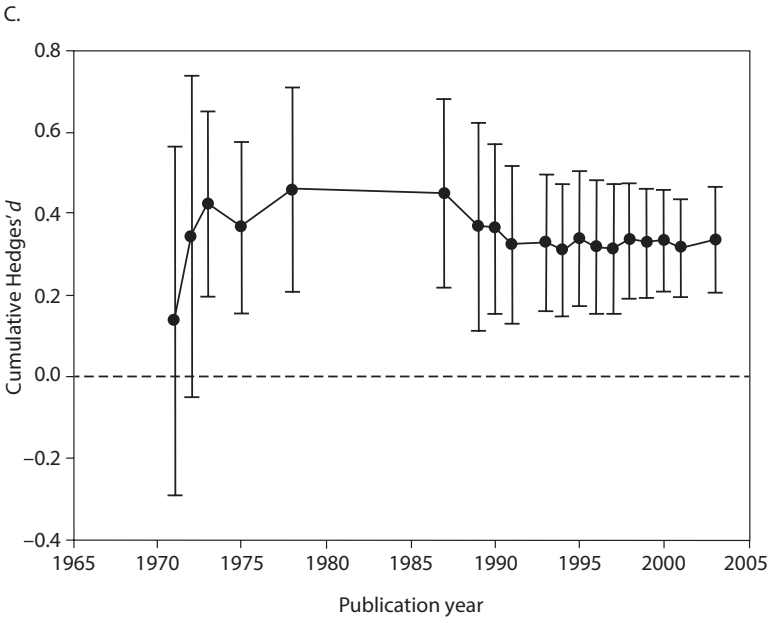


Figure 15.2. Scatterplot (*A*), cumulative meta-analysis plot (*B*), and cumulative meta-analysis trajectory (*C*) of studies examining effects of male mating history on female reproductive output (data from Torres-Vila and Jennions 2005). CMA plot was plotted using Comprehensive Meta-Analysis software.



expected by chance ($z = -0.89$, $P < 0.05$). Therefore, effects of male mating history on female reproductive output do not exhibit pronounced temporal changes, indicating that the conclusions of the meta-analysis are stable. Nearly the same results would be obtained if the meta-analysis was conducted in the early 1990s instead of 2003.

CONCLUSIONS AND BEST PRACTICE RECOMMENDATIONS

Significant temporal changes in the magnitude of effect sizes appear to be common in ecology and evolutionary biology as well as in other scientific disciplines. They often lead to changes in statistical significance of the mean effect over time and present a fundamental problem for research synthesis if meta-analyses conducted at different points in time provide different conclusions. If the effects decrease with time (as often appears to be the case in ecology), then early meta-analyses are likely to overestimate the effect. This is of particular concern in conservation biology and applied ecology because it is hoped that the use of meta-analysis in these fields will facilitate communication of research findings to policy makers and lead to the development of evidence-based management policies (Chapter 26).

We recommend that ecologists and evolutionary biologists always explore temporal patterns in effect sizes when conducting a meta-analysis; particularly when the total heterogeneity of effect sizes is significant and the data set offers a sufficient temporal span (at least 10 years). If graphical methods (scatterplots or CMA plots) indicate temporal trends, the extent of the temporal changes in effect sizes should be tested by using the statistical methods (e.g., regression analysis) described above.

Examination of temporal trends in effect sizes should become a routine procedure in ecological and evolutionary meta-analyses, just as tests for publication bias currently are (Chapter 14). Scrutinizing temporal trends is useful for several reasons. First, exploration of temporal trends in meta-analytic data sets is crucial for assessment of stability of the results and the

sufficiency of the data. Second, temporal changes in effect size are often indicative of other sources of heterogeneity, such as publication bias or heterogeneity among studies with respect to research methods or the study organisms. Examination of temporal trends in effect sizes is thus a good diagnostic tool for detection of sources of heterogeneity. It might also allow early detection of the point in time when the evidence is sufficient and stable enough to provide the basis for management recommendations. Timely detection of temporal changes in effect sizes that might indicate the need for changes in previously accepted management policies is also possible. This could ultimately result in a significant savings of time and resources in the development of management strategies, and would result in more effective conservation actions. The lack of stability in effect sizes over time could also be used as justification for the need for more research on the topic.

We also recommend that, as meta-analysis becomes more common in ecology and evolutionary biology, biologists adopt the practice of updating meta-analyses on the same topic when a sufficient number of new studies becomes available. In medicine, regular, biennial updating of research evidence by meta-analyses is a standard practice in initiatives like the Cochrane Collaboration (Higgins and Green 2011). In contrast, there is no similar procedure yet in ecology and evolutionary biology and researchers are often discouraged from repeating a review, largely because the majority of journals strongly emphasize novelty as a major criterion for publication (Palmer 2000).

Cumulative meta-analysis represents a useful tool for updating summary results as evidence accumulates. In terms of sufficiency and stability of data, the CMA plot would serve as an indicator of the robustness of the conclusions drawn from the analysis (Mullen et al. 2001). If the cumulative effect size and the confidence interval show no evidence of stabilization, the results of the analysis should be interpreted with caution and potential causes of the temporal changes in effect sizes should then be examined. However, because of the limitations of the CMA as a statistical procedure, we recommend that it is always complemented by the formal statistical analysis of the temporal trend.