Meta Analyses in Studying Insect Declines Technical Review Plan

**Intro – what meta-analysis is and why we need it, what is the benefit of doing it**

Gurevitch, 2018 - quantitative, scientific synthesis of research results. Statistical procedure. Term first introduced in 1970s. First used by ecologists some 25 years ago. Resolve contradictory research outcomes. Reach broad generalisations. Make more comprehensive picture. Individual primary study may now be seen as contribution towards accumulation of knowledge rather than revealing conclusive answer. Systematic reviews can be combined with quantitative meta-analyses to assess the magnitude of the outcome across relevant primary studies and to analyse the causes of variation. Systematic reviews aim to be transparent, reproducible and updatable, and to address well-defined questions.

Arnqvist, 1995 – generate higher order conclusions. Summarise and analyse multiple independent studies, as studies typically differ in the magnitude of effects and in the occurrence of significant results.

Nakagawa and Santos, 2012 – MA techniques originally developed in medical and social sciences. Biological MA often integrates complex data composed of multiple strata.

Cote and Reynolds, 2012 - geographically or taxonomically broad-scale comparisons.

Gurevitch and Hedges, 1999 – how large is overall effect? How consistent across studies? Are difference due to meaningful categories?

Vetter, 2013 – higher statistical power and better precision, assess broader scope. Part of systematic review, which involves clearly formulated research question, an extensive literature search. Quantitative synthesis of the data (normally by a meta-analysis), and interpretation of the results. MA allows us to calculate the magnitude rather just than the existence of an effect AND assess if effect sizes are homogeneous across studies. If variation very high, meta-regression can help explain.

Stewart, 2010 – for evidence-based decision making. MA illustrate danger of over-reliance on info from limited range of studies. Ecological MA require focus on exploration of heterogeneity in almost all case due to studies being measured on different spatio-temporal scales. Only then can generalisability be assessed.

Lortie, 2015 – MA should include transparency, replicability, and clear statement of purpose. Allows testability, generality, consistency, accuracy, and bias in the studies published on a topic at the time of analysis. Does not test hypotheses but is a means to explore the strength of evidence associated with hypotheses.

Koricheva and Gurevitch, 2014 – MA allows testing for covariates which are logistically difficult or impossible to test within a single empirical study. Reduce the risk of single experiments capturing media attention and inappropriately alarming or falsely comforting the public and policymakers. Can help identify knowledge gaps.

Nakagawa and Poulin, 2012 – main role of review piece of any kind is to provide up to date overview of the state of knowledge.

*Why better than other methods?*

Gurevitch, 2018 - Better than narrative review which cannot accurately summarize results across studies.

Gurevitch, 2018 – better than vote-counting (count up number of significant and non-significant results) which is statistically flawed. Outcomes can be very misleading. Technique long since been discredited.

Arnqvist, 1995 - less subjective than narrative reviews, based on a formal, predetermined set of statistical procedures rather than individual interpretations.

Gurevitch and Hedges, 1999 - vote counting = low statistical power, and most importantly it fails to provide critical information on the overall results.

Vetter, 2013 - systematic review differs substantially from a narrative review in its transparency and replicability. Vote-counting - statistically problematic

Harrison, 2011 – ameliorates problems with vote-counting - effect sizes more informative, they also represent continuous variables that can be combined and compared. More clearly needs driven and evidence based.

Nakagawa and Poulin, 2012 – outperforms qualitative reviews, can even generate new hypotheses and future research directions. Narrative review cannot be used to test a hypothesis. Vote-counting ignores study quality in terms of sample sizes and methodologies, often leading to erroneous conclusions. MA test hypothesis in rigorous a way as empirical studies. Have ability to detect small effects which any single study cannot reliably detect.

Cadotte, 2012 – analysis of 240 MA in ecology. Publication rates increased through time, becoming more comprehensive, including more studies and greater temporal range of studies (due to increased data availability and accessibility). MA have power to detect significant differences in the pooled data even when individual datasets fail to detect significant effects. Can use data that have limited geographical extent to search for broader generalities. Can definitively show that certain patterns are generalizable.

*Overview of process of MA:*

Gurevitch, 2018 – One or more outcomes in the form of effect sizes are extracted from each study. Put outcomes of different studies on same scale. Effect sizes entered into a statistical model. Use weighting based on precision – larger studies with higher precision weighted more heavily.

Nakagawa, 2017 - Moderators are equivalent to explanatory (independent) variables or predictors in a normal linear model. Models that examine the effects of moderators are referred to as meta-regressions.

Nakagawa and Santos, 2012 – combines common effect-size statistics by accounting for the sample sizes of the studies. Next to finding general trend, the next most important function is to quantify heterogeneity. Do this with Cochran's Q and I2. I2 becoming increasingly popular. I2 = 25, 50 and 75% are considered as low, moderate and high heterogeneity, respectively.

Gurevitch and Hedges, 1999 – choosing effect size statistic, calculate grand mean effect size, as well as means/slopes for different explanatory variables, determine confidence intervals around these, determine consistency within and among categories of studies.

*R packages:*

Gurevitch, 2018 – metafor released in 2010.

Viechtbauer, 2010 – made metafor - fitting the meta-analytic fixed- and random-effects models and allows for the inclusion of moderators variables. Escalc() function for effect size estimates. Fit models with rma.uni().

Lajuenesse, 2016 – METAGEAR – there are few resources available to facilitate research synthesis as a whole. Can be used to sieve bibliographic information. Can automate the download of these studies. Can generate and update flow diagrams. Can extract data from figures. Previously ecologists relied heavily on METAWIN (no longer maintained), but METAFOR now dominates.

**Methods used**

*Effect size*

Arnqvist, 1995 - standardized difference between means of experimental and control groups or the Pearson product moment correlation coefficient.

Viechtbauer, 2010 - odds ratio, relative risk, risk difference, the correlation coefficient, and the (standardized) mean difference

Nakagawa, 2017 – difference between the means of two groups – use standardized mean difference (Cohen's d or Hedges' g) or natural log of the response ratio.

* These will be the 2 most relevant to ecology

Nakagawa and Santos, 2012 - Standardized mean difference was the most popular category of effect-size statistics reported. Response ratio was the most commonly used effect-size statistic overall, although almost exclusive to the field of ecology and evolution. Nature of meta-analysis usually dictates which effect-size statistic is used, but not always straightforward.

Hedges, 1999 - response ratio (the ratio of mean outcome in the experimental group to that in the control group). Argue it is better for ecological MA than standardised difference between means, which is not always meaningful way to summarise experiments. Ratio quantifies proportional change. Divide mean of experimental group by mean of control group to get response ratio. Do natural log of this to linearise metric and make distribution more normal in small samples. Can't use if substantial proportion of studies have zero as their control mean.

Koricheva and Gurevitch, 2014 – plant ecology mostly uses response ratios and standardized mean differences. Effect sizes based on binary data are seldom used in plant ecology because most variables are continuous rather than binary.

Harrison, 2011 - measure rates of change in independent groups - response ratio can be used, generally log-transformed to linearize and normalize the raw ratios. Binary response data e.g. nest success or survival – use risk ratio or odds ratio. BUT this is rarely used in ecology.

*Types of model*

Gurevitch, 2018 – Fixed effect model = assumes variation in effect sizes among studies is due to within-study (sampling) variance and that all studies share common "true" effect. Vs random effects model = in addition to sampling variance, assumes true effects from different studies also differ from one another. So random-effects models have extra variance component to account for between-study variance (heterogeneity). To identify magnitude and sources of variation, use meta-regression.

Nakagawa and Santos, 2012 - random-effects meta-analysis uses the more reasonable assumption that each study has a ‘true’ effect size different from each other. Once heterogeneity is identified, the next step in meta-analysis is to incorporate moderators which may explain the observed heterogeneity. In other words, we move onto meta-regression (also called mixed-effects meta-analysis). Because heterogeneity is almost always expected, we suggest that meta-regression should be the default meta-analytic model for biological meta-analysis. Can use AIC/DIC to select better fitting meta-regression models.

* Heterogeneity especially true for insect trends

Gurevitch and Hedges, 1999 – fixed effect - share a common, ‘‘true’’ effect size, and estimates differ from one another by sampling error only. Vs random effects - true effect size is expected to differ among studies, and the goal of the analysis is to quantify the variation in the effect parameters. More appealing in ecology. Mixed-effect - effects of the groupings of experiments on effect parameters are fixed effects, within groupings are taken to be random effects (i.e. random-effects model with moderators).

Noble, 2017 - total heterogeneity in ecological and evolutionary meta-analyses to be very high on average (~92%), which indicates that random-effects models are more appropriate for typical meta-analyses in ecology and evolution. Multilevel meta-regression models have developed rapidly and can now deal with many sources of nonindependence, while allowing one to explore moderator variables to test a rich set of methodological and biological hypotheses explaining variation among effect sizes.

Borenstein, 2010 – fixed effect – one true effect size that underlies all the studies, different in observed effects due to sampling error (within study variance). One source of variance. Would be better to call these common effect models. Vs random effect - allow the true effect sizes to differ, effect size varies from study to study. 2 sources of variance.

Under the random-effects model the confidence interval will always be wider and the weights will always be more similar to each other than under the fixed-effect model.

Random-effects model yields a less precise estimate of the combined effect than does the fixed-effect model (as more sources of variation), some find it surprising that the random-effects model also yields a different estimate of the combined effect itself. Due to weights being different under the 2 models.

Random-effects - estimate the mean of a distribution of effects, so want to be sure that all these effect sizes are represented in the summary estimate. Cannot discount a small study by giving it a very small weight, cannot give too much weight to a very large study. Random-effects weighting based on within-study variance plus a constant. This constant reduces relative differences among the weights, so weights assigned are more balanced.

The strategy of starting with a fixed-effect model and then moving to a random-effects model if the test for heterogeneity is significant relies on a flawed logic and should be strongly discouraged.

Harrison, 2011 – if effect sizes shows more heterogeneity than would be expected due to sampling error alone by calculating the Q statistic, re-calculate the inverse variance weights in order to calculate a random-effects version. Or use mixed-effects

* Borenstein, 2010 says don't do fixed effect first and then move to random if seems appropriate – think about it first

Assink and Wibbelink, 2016 – fitting three-level meta-analytic models. Strong method for dealing with dependency of effect sizes. Relatively unknown. Use rma.mv function in metafor. Can be extended to include moderators (3 level mixed effects model). If non-independent, there is overlap in info so it is inflated leading to overconfidence in results. Using three level (multi-level) MA, can extract all relevant effect sizes from each primary study without needing to reduce the number of effect sizes - all information can be preserved and maximum statistical power can be achieved. Considers three different variance components distributed over the three levels of the model: sampling variance of the extracted effect sizes at level 1; variance between effect sizes extracted from the same study at level 2; and variance between studies at level 3. In short, this model allows effect sizes to vary between participants (level 1), outcomes (level 2), and studies (level 3).

*Weighting*

Gurevitch, 2018 – importance of weighting. Unweighted analyses increase the influence of small studies. Nevertheless, the variances needed for weighted MA are frequently unavailable due to poor reporting.

Arnqvist, 1995 – No two studies in a set of studies are equally ‘reliable’ - accounted for by giving estimates from different studies different weights, primarily based on their sample size.

Gurevitch and Hedges, 1999 – weight effect sizes for statistical analysis by the inverse of the sampling variance of the effect size. Counting large studies more heavily than small ones, which often seems reasonable in summarizing overall results.

Hedges, 1999 - differ in precision (standard error) – so weighting of the individual study estimates giving greater weight to experiments whose estimates have greater statistical precision (smaller standard error) will increase the precision of the combined estimate.

*Assessing heterogeneity*

Senior, 2016 - heterogeneity will be present in ecological/evolutionary meta-analyses due to the system-specific nature of biological phenomena. Reviewed 700 studies, total heterogeneity was reported in fewer than 40%. Estimates are well above “high” heterogeneity (i.e., 75%), based on widely adopted benchmarks (in medical fields). Due to heterogeneity not being reported when low? BUT even studies that had significant meta-regression results did not report heterogeneity.

*Reporting*

Vetter, 2013 – forest plot allows reader to quickly assess the number of studies that form the summary effect, the precision of the included studies and the homogeneity/heterogeneity across effect sizes.

Anzures-Cabrera, 2010 – forest plot – displays effect estimates and their confidence intervals. Area of the block is proportional to the weight assigned to that study. May include the result of the overall effect from a meta-analysis, often using a diamond to distinguish it from the individual studies. Funnel plot – effect estimate against a measure of precision. To detect bias. Asymmetric funnel indicates a relationship between effect size and precision = publication bias. Galbraith plots appropriate when there are more studies than can comfortably be displayed on a forest plot. L’Abbé plots appropriate only for studies comparing two groups. Display meta-regression with simple scatter plot. Forest plot best and should continue to be first choice when viable.

Harrison, 2011 – funnel plot of effect size vs. study size to assess publication bias. Effect sizes reported in a number of studies should be symmetrically distributed around the underlying true effect size, with more variation from this value in smaller studies than in larger ones.

Page, 2021 – PRISMA 2020 statement. Help reviewers transparently report why the review was done, what the authors did, and what they found. 27 items recommended for reporting. Designed primarily for systematic reviews that evaluate effect of health interventions, but can be applied more widely. Allows reader to assess the appropriateness of the methods, and therefore the trustworthiness of the findings.

**Limitations**

*Poor reporting/understanding/mis-use*

Gurevitch, 2018 – editors and authors are not fully aware of well-developed meta-analytic methodology. Should only be applied to studies that use well-established statistical procedures. Term often misapplied, regardless of the rigour of the methodology. the variances needed for weighted MA are frequently unavailable due to poor reporting. Could introduce pre-registration to overcome this?

Nakagawa, 2017 - Documentation on keyword strings and inclusion criteria is often also very poor, making replication of search outcomes difficult or impossible.

Cote and Reynolds, 2012 – looked at MA linking conservation to evolution. Meta-analysts typically had to drop over half of the papers they wanted to summarize because of lack of basic information such as variances and sample sizes.

Gurevitch and Hedges, 1999 - Ecological experiments commonly fail to report sample sizes and variances.

Vetter, 2013 – MA not well defined in ecology and conservation. Vague and inconsistent utilisation. Of 133 evaluated articles, one 1 fulfilled all requisite steps in both rounds of requirement screening. Worrying that most meta-analysts do not consider or even mention the heterogeneity/variability of effect sizes given high heterogeneity in natural systems. Want to avoid this to avoid misinterpretation among scientists, public, and decision makers. Realise they couldn't address the quality of raw data.

Koricheva and Gurevitch, 2014 – confusion between MA and vote-counting. In plant ecology (26%) studies assessed used unweighted meta-analyses, largely because they reported that estimates of variance (standard deviations and standard errors) were not available in primary studies. 33% do not specify the statistical model used. 61% did not include any tests for publication bias.

Nakagawa and Poulin, 2012 – even though you are able to detect publication bias with MA, ecologists often don't perform the analyses required. Could be resolved by journals imposing clear guidelines of statistical reporting. PRISMA statement developed for health sciences can potentially be adopted for MA in ecology and evolution and should be used.

*Publication bias*

Gurevitch, 2018 – publication bias invalidates the use of MA

Arnqvist, 1995 - bias that will result when the studies included in the MA are not representative of all studies conducted. Due to biases either in publication rates or in selection/retrieval of studies.

Nakagawa, 2017 – have to assume no publication bias, which is unlikely. Should check using statistical and graphical tools e.g. funnel plots, Egger's regression tests.

Nakagawa and Santos, 2012 - statistically significant results are more likely to be published. BUT in MA, we have the ability to detect publication bias. Also means we will spur further investigations. Funnel plots (not statistical test, assess funnel asymmetry) and regression tests (statistical test) to identify existence, trim and fill method to assess impact and adjust for impact by restoring funnel symmetry by using existing data points to impute missing studies. BUT this is designed for MA where independence can be assumed. Publication bias is only one of the potential causes of funnel asymmetry, it could also just be down to chance.

Gurevitch and Hedges, 1999 – detecting publication bias is based on examination of the relation between standard error and effect size. BUT, relation between sample size and effect size may reflect rational experimental design rather than publication bias.

Stewart, 2010 – over-estimate effect size due to publication bias. Could introduce global register, which require submission of objectives and methods prior to commencement of data collection.

Duval and Tweedie, 2000 – trim and fill method for accounting for publication bias – formalises the use of funnel plots to estimate and adjust for missing studies.

*Non-independence*

Nakagawa, 2017 - Failing to account for non-independence among effect sizes can lead to erroneous conclusions.

Nakagawa and Santos, 2012 – Non-independence can arise from observations from related sources, or phylogenetic relatedness of species. Random effects model designed for meta-analyses where the number of effect sizes equals the number of studies. Biological meta-analyses will frequently include studies that often provide more than one effect-size estimate e.g. assessing one or more functions by multiple insect species, or ii) repeated measures of one or more functions by the same species. Some meta-analyses opted to include only one effect-size estimate per study, losing statistical power and potentially forgoing important information. To address this issue most effectively, we should use multilevel/hierarchical models, which explicitly model correlations within the different levels.

Non-independence (different effect sizes may be correlated because the data on which they are based are correlated) can lead to underestimates of the standard error of the mean effect and therefore liberal evaluations of the statistical significance of effects.

- i.e. type 1 error – false positive

Could conduct a different meta-analysis for each kind of effect measure OR better, use of all of the data via multivariate methods that explicitly model the dependence structure.

Noble, 2017 - Nonindependence will artificially inflate sample size, increasing the magnitude of the denominator in variance equations and, thus, decreasing the sampling error variance for an effect size. Log response ratio effect size is less susceptible to non-independence than standardized mean difference.

*Other*

Gurevitch, 2018 – ecological studies are too heterogenous to be combined statistically in a meaningful way, ecology is best served by accumulating a catalogue of case studies.

Gurevitch, 2018 – can highlight areas in which evidence is deficient, cannot overcome these deficiencies.

Gurevitch, 2018 – Individual primary study may now be seen as contribution towards accumulation of knowledge rather than revealing conclusive answer. De-value primary studies?

Stewart, 2010 - all syntheses are constrained by the quality of available data and the standards of reporting in primary research.

Nakagawa and Poulin, 2012 - not suitable for appraising new hypotheses, for which no or few empirical studies exist.

**E.g. of meta-analyses on insect declines**

Van Klink, 2020 - 166 long-term surveys of insect assemblages. 41 countries. 1925 to 2018. Average decline of terrestrial insect abundance by ~9% per decade and an increase of freshwater insect abundance by ~11% per decade. Analysed the data using a hierarchical Bayesian model accounting for variation at the study, study area, and site level. Mixed effect models. Used temp and precipitation as moderators but found no associations. Data not representatively spread across the world. Trends in protected areas weaker than those in unprotected – suggests possible association between insect trends and land-use

* Though this last bit not properly tested
* Also criticised by Jahnig, 2021 – abundance and biomass poor indicator of status – sensitive species may be being replaced by tolerant ones. Also, poorly represent global trends.

Dirzo, 2014 - Lepidopteran species richness is on average 7.6 times higher in undisturbed than disturbed sites, and total abundance is 1.6 times greater. Used Hedge's g effect size. Used funnel plots, trim and fill analysis, forest plots. Calculated 52 diversity effect sizes from 15 studies.

**The future of meta-analyses**

Gurevitch, 2018 - use of machine learning and artificial intelligence (AI) to screen studies for inclusion

Lajeunesse, 2016 - automating PDF annotations, text-mining tools to enhance data retrieval and extraction, methods to estimate statistical power, multiple-imputation tools, and better approaches to assess publication bias.

**Other**

Nakagawa, 2017 - Meta-analyses should focus on biological importance rather than on p values.

Nakagawa, 2017 - New insights from meta-analyses are known as ‘review-generated evidence’ (as opposed to ‘study-generated evidence’)

Nakagawa, 2017 - updating is still rare in biological meta-analyses

Hedges, 1999 - If the between-experiment variation is too large, one might question whether the experimental results are similar enough to warrant combination.

Vetter, 2013 – assessed quality of MA – based rating on the classical frequentist approach, which was far more common among the evaluated articles.

Stewart, 2010 - Ecological meta-analysts have recognized the benefit of using hierarchical Bayesian models to explore complex data, but their full potential remains untapped.

Lortie, 2015 – on-going shift away from p values

Koricheva and Gurevitch, 2014 - controversial whether one can use research syntheses to make causal inferences

Koricheva and Gurevitch, 2014 – QC – need details on:

* Effect sizes + weighted
* Bibliographic search + details of all studies
* Study inclusion/exclusion criteria
* Non-independence considered
* Statistical model described
* Heterogeneity quantified and explored for causes
* Publication bias
* Temporal changes in effect size?
* Sensitivity analyses to test robustness of results – would results differ if MA was conducted in different year?
* Data set included as appendix
* Include PRISMA flow diagram - shows how studies used in meta-analysis were identified, screened and assessed for eligibility

Borenstein, 2010 – if number of studies is very small, can perform Bayesian meta-analysis, where to estimate of between-study variance is based on data outside of the current set of studies.

Anzures-Cabera, 2010 - heterogeneity variance in a random-effects meta-analysis is imprecisely estimated when the number of studies is not large.

Harrison, 2011 - Errors or omissions made at the planning stage can create weeks of extra work. Far preferable to produce a data base which includes information that you later decide not to use. Meta-analyses are only as good as the data used.

My understanding of different types of models:

* Fixed effect models – studies differ due to sampling error
* Random effects models - studies differ due to sampling error and between study variation
* Mixed effect models / meta-regression – include moderator
* Multiple meta-regression – includes several predictor variables/moderators, and their interaction
* Multilevel models – takes non-independence into account – built into the structure of the model