Stage 1 Registered Report

Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology

Tim Parker, Department of Biology, Whitman College, Walla Walla WA, 99362 USA
Hannah Fraser, School of BioSciences, University of Melbourne, Melbourne, Victoria, Australia
Shinichi Nakagawa, Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Randwick, New South Wales, Australia
Elise Gould, School of BioSciences, University of Melbourne, Melbourne, Victoria, Australia
Simon Griffith, Department of Biological Sciences, Macquarie University, Sydney, New South Wales 2109, Australia
Pete Vesk, School of BioSciences, University of Melbourne, Melbourne, Victoria, Australia
Fiona Fidler, School of BioSciences and School of Historical and Philosophical Studies, University of Melbourne, Melbourne, Victoria, Australia
and all recruited analysts and analysis reviewers

Abstract

Although variation in effect sizes and predicted values among studies of similar phenomena is inevitable, there is evidence that such variation may far exceed what might be produced by sampling error. This evidence comes from a growing meta-research agenda that seeks to describe and explain variation in reliability of scientific results. One possible explanation for variation among results is differences among researchers in the decisions they make regarding statistical analyses. The best evidence for this comes from a recent social science study that asked 29 different research teams to answer the same question independently by analyzing the same data set. Although many of the effect sizes were similar, some differed substantially from the average. We plan to implement an analogous study in ecology and evolutionary biology, a field in which there has been no empirical exploration of the variation in effect sizes or model predictions of dependent variables generated by analytical decisions of different researchers. We have obtained two unpublished data sets, one from evolutionary ecology and one from conservation ecology, and we will recruit as many independent scientists as possible to conduct analyses of these data to answer prespecified research questions. We will also recruit peer reviewers to rate the analyses based on their methodological descriptions so that we have multiple ratings of each analysis. Next we will quantify the variability in choices of independent variables among analyses and, using meta-analytic techniques, describe and quantify the degree of variability among effect sizes and predicted values for each of the data sets. Finally, we will quantify the extent to which deviation of individual effect sizes and predicted values from the meta-analytic mean for that data set is explained by peer review ratings and by the 'uniqueness' of the set of variables chosen for the analysis by each team.

Keywords

credibility revolution, heterogeneity, meta-analysis, meta-science, replicability, reproducibility

Introduction

One value of science derives from its production of replicable, and thus reliable, results. When we repeat a study using the original methods we should be able to expect a similar result. However, perfect replicability is not a reasonable goal. Effect sizes will vary, and even reverse in sign, by chance alone (Gelman and Weakliem 2009). However, observed patterns can differ for other reasons as well. It could be that we do not sufficiently understand the conditions that led to the original result so when we seek to replicate it, the conditions differ due to some 'hidden moderator'. This hidden moderator hypothesis is described by meta-analysts in ecology and evolutionary biology as 'true biological heterogeneity' (Senior et al. 2016). This idea of true heterogeneity is popular in ecology and evolutionary biology, and there are good reasons to expect it to be true in the complex systems in which we work (Shavit and Ellison 2017). However, despite similar expectations in psychology, recent evidence in that discipline contradicts the hypothesis that moderators are common obstacles to replicability, as variability in results in a large 'many labs' collaboration was largely unrelated to commonly hypothesized moderators such as the conditions under which the studies were administered (Klein et al. 2018). Another possible explanation for variation in effect sizes is that researchers often present biased samples of results, thus reducing the likelihood that later studies will produce similar effect sizes (Open_Science_Collaboration 2015, Parker et al. 2016, Forstmeier et al. 2017, Fraser et al. 2018). It also may be that although researchers did successfully replicate the conditions, the experiment, and measured variables, analytical decisions differed sufficiently among studies to create divergent results (Simonsohn et al. 2015, Silberzahn et al. 2018).

Analytical decisions vary among studies because researchers have many options. Researchers need to decide how to exclude possibly anomalous or unreliable data, how to construct variables, which variables to include in their models, and which statistical methods to use. Depending on the data set, this short list of choices could encompass thousands or millions of possible alternative specifications (Simonsohn et al. 2015). However, researchers making these decisions presumably do so with the goal of doing the best possible analysis, or at least the best analysis within their current skill set. Thus it seems likely that some specification options are more probable than others, possibly because they have previously been shown (or claimed) to be better, or because they are more well known. Of course, some of these different analyses (maybe many of them) may be equally valid alternatives. Regardless, on probably any topic in ecology and evolutionary biology, we can encounter differences in choices of data analysis. The extent of these differences in analyses and the degree to which these differences influence the outcomes of analyses and therefore studies' conclusions are important empirical questions. These questions are especially important given that many papers draw conclusions after applying a single method, or even a single statistical model, to analyze a data set.

The possibility that different analytical choices could lead to different outcomes has long been recognized, and various efforts to address this possibility have been pursued in the literature. For instance, one common method in ecology and evolutionary biology involves creating a set of candidate models, each consisting of a different (though often similar) set of predictor variables, and then, for the predictor variable of interest, averaging the slope across all models (i.e. model averaging) (Burnham and

Anderson 2002, Grueber et al. 2011). This method reduces the chance that a conclusion is contingent upon a single model specification, though use and interpretation of this method is not without challenges (Grueber et al. 2011). Further, the models compared to each other typically differ only in the inclusion or exclusion of certain predictor variables and not in other important ways, such as methods of parameter estimation. More explicit examination of outcomes of differences in model structure, model type, data exclusion, or other analytical choices can be implemented through sensitivity analyses (e.g., Noble et al. 2017). Sensitivity analyses, however, are typically rather narrow in scope, and are designed to assess the sensitivity of analytical outcomes to a particular analytical choice rather than to a large universe of choices. Recently, however, analysts in the social sciences have proposed extremely thorough sensitivity analysis, termed 'multiverse analysis' (Steegen et al. 2016) or the 'specification curve' (Simonsohn et al. 2015), as a means of increasing the reliability of results. With these methods, researchers identify relevant decision points encountered during analysis and conduct the analysis many times to incorporate all plausible decisions made at each of these points. The study's conclusions are then based on a broad set of the possible analyses and so allow the analyst to distinguish between robust conclusions and those that are highly contingent on particular model specifications. These are useful outcomes, but specifying a universe of possible modelling decisions is not a trivial undertaking. Further, the analyst's knowledge and biases will influence decisions about the boundaries of that universe, and so there will always be room for disagreement among analysts about what to include. Including more specifications is not necessarily better. Some analytical decisions are better justified than others, and including biologically implausible specifications may undermine this process. Regardless, these powerful methods have yet to be adopted, and even more limited forms of sensitivity analyses are not particularly widespread. Most studies publish a small set of analyses and so the existing literature does not provide much insight into the degree to which published results are contingent on analytical decisions.

Despite the potential major impacts of analytical decisions on variance in results, the outcomes of different individuals' data analysis choices have received limited empirical attention. The only formal exploration of this that we are aware of were (1) an analysis in social science that asked whether male professional football (soccer) players with darker skin tone were more likely to be issued red cards (ejection from the game for rule violation) than players with lighter skin tone (Silberzahn et al. 2018) and (2) an analysis in neuroimaging which evaluated nine separate hypotheses involving the neurological responses detected with fMRI in 108 participants divided between two treatments in a decision making task (Botvinik-Nezer et al. 2019).

In the red card study, twenty-nine teams designed and implemented analyses of a data set provided by the study coordinators (Silberzahn et al. 2018). Analyses were peer-reviewed (results blind) by at least two other participating analysts; a level of scrutiny consistent with standard pre-publication peer-review. Among the final 29 analyses, odds-ratios varied from 0.89 to 2.93, meaning point estimates varied from having players with lighter skin tones receive more red cards (odds ratio < 1) to a strong effect of players with darker skin tones receiving more red cards (odds ratio > 1). Twenty of the 29 teams found a statistically-significant effect in the predicted direction of players with darker skin tones being issued more red cards. This degree of variation in peer-reviewed analyses from identical data is striking, but the generality of this finding has only just begun to be formally investigated.

In the neuroimaging study, 70 teams evaluated each of the nine different hypotheses with the available fMRI data (Botvinik-Nezer et al. 2019). These 70 analysts followed a divergent set of workflows that produced a wide range of results. The rate of reporting of statistically significant support for the nine hypotheses ranged from 21% to 84%, and for each hypothesis on average, 20% of research teams observed effects that differed substantially from the majority of other teams. Some of the variability in results among studies could be explained by analytical decisions such as choice of software package, smoothing function, and parametric versus non-parametric corrections for multiple comparisons. However, substantial variability among analyses remained unexplained, and presumably emerged from the many different decisions each analyst made in their long workflows. Such variability in results among analyses from this data set and from the very different red-card data set suggests that sensitivity of analytical outcome to analytical choices may characterize many distinct fields.

To further develop the empirical understanding of the effects of analytical decisions on study outcomes, we chose to estimate the extent to which researchers' data analysis choices drive differences in effect sizes, model predictions, and qualitative conclusions in ecology and evolutionary biology. This is an important extension of the meta-research agenda of evaluating factors influencing replicability in ecology, evolutionary biology, and beyond (Fidler et al. 2017). To examine the effects of analytical decisions, we have two different data sets and we will recruit researchers to analyze one or the other of these data sets to answer a question we have defined. The first question is "To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?" To answer this question, we have a data set that includes brood size manipulations from 332 broods conducted over three years at Wytham Wood, UK. The second question is "How does grass cover influence *Eucalyptus spp.* seedling recruitment?" For this question, analysts will use a data set that includes, among other variables, number of seedlings in different size classes, percentage cover of different life forms, tree canopy cover, and distance from canopy edge from 351 quadrats spread among 18 sites in Victoria, Australia.

We will explore the impacts of data analysts' choices with descriptive statistics and with a series of tests to attempt to explain the variation among effect sizes and predicted values of the dependent variable produced by the different analysis teams for both data sets separately. To describe the variability, we will present forest plots of the standardized effect sizes and predicted values produced by each of the analysis teams, estimate heterogeneity (both absolute [τ^2] and proportional [I^2]) in effect size and predicted values among the results produced by these different teams, and calculate a similarity index that quantifies variability among the predictor variables selected for the different statistical models constructed by the different analysis teams. These descriptive statistics will provide the first estimates of the extent to which explanatory statistical models and their outcomes in ecology and evolutionary biology vary based on the decisions of different data analysts. We will then quantify the degree to which the variability in effect size and predicted values can be explained by (1) variation in the quality of analyses as rated by peer reviewers and (2) the similarity of the choices of predictor variables between individual analyses.

Methods

This project involves a series of steps (1-6) that begin with identifying data sets for analyses and continue through recruiting independent groups of scientists to analyze the data, allowing the scientists to analyze the data as they see fit, generating peer review ratings of the analyses (based on methods, not results), evaluating the variation in effects among the different analyses, and producing the final manuscript. We estimate that this process, from the time of an in-principle acceptance of this stage 1 Registered Report, will take 11 months (Table 1). The factor most likely to delay our timeline is the rate of completion of the original set of analyses by independent groups of scientists.

Step 1: Select Datasets

We will use two previously unpublished datasets, one from evolutionary biology and the other from ecology and conservation.

Evolutionary ecology

Our evolutionary ecology data set is relevant to a sub-discipline of life-history research which focuses on identifying costs and trade-offs associated with different phenotypic conditions. These data were derived from a brood-size manipulation experiment imposed on wild birds nesting in boxes provided by researchers in an intensively studied population. Understanding how the growth of nestlings is influenced by the numbers of siblings in the nest can give researchers insights into factors such as the evolution of clutch size, determination of provisioning rates by parents, and optimal levels of sibling competition (Vander Werf 1992, DeKogel 1997, Royle et al. 1999, Verhulst et al. 2006, Nicolaus et al. 2009). Data analysts will be provided this data set and instructed to answer the following question: “To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?”

Researchers conducted brood size manipulations and population monitoring of blue tits (*Cyanistes caeruleus*) at Wytham Wood, a 380 ha woodland in Oxfordshire, U.K (1º 20’W, 51º 47’N). Researchers regularly checked approximately 1100 artificial nest boxes at the site and monitored the 330 to 450 blue tit pairs occupying those boxes in 2001-2003 during the experiments. Nearly all birds made only one breeding attempt during the April to June study period in a given year. At each blue tit nest, researchers recorded the date the first egg appeared, clutch size, and hatching date. For all chicks alive at age 14 days, researchers measured mass and tarsus length and fitted a uniquely numbered, British Trust for Ornithology (BTO) aluminium leg ring. Researchers attempted to capture all adults at their nests between day 6 and day 14 of the chick-rearing period. For these captured adults, researchers measured mass, tarsus length, and wing length and fitted a uniquely numbered BTO (British Trust for Ornithology) leg ring. During the 2001-2003 breeding seasons, researchers manipulated brood sizes using cross fostering. They matched broods for hatching date and brood size and moved chicks between these paired nests one or two days after hatching. They sought to either enlarge or reduce all manipulated broods by approximately one fourth. To control for effects of being moved, each reduced brood had a portion of its brood replaced by chicks from the paired increased brood, and vice versa. Net manipulations varied from plus or minus four chicks in broods of 12 to 16 to plus or minus one chick in broods of 4 or 5. Researchers left approximately one third of all broods unmanipulated. These unmanipulated broods were not selected systematically to match manipulated broods in clutch size or laying date. We have mass and tarsus length data from 3720 individual chicks divided among 167 experimentally enlarged broods, 165 experimentally reduced broods, and 120 unmanipulated broods. The full list of variables included in the data set is available in Table S1 in the supplementary materials.

Ecology and conservation

Our ecology and conservation data set is relevant to a sub-discipline of conservation research which focuses on investigating how best to revegetate private land in agricultural landscapes. These data were collected on private land under the Bush Returns program, an incentive system where participants entered into a contract with the Goulburn Broken Catchment Management Authority and received annual payments if they executed predetermined restoration activities. This particular data set is based on a passive regeneration initiative, where livestock grazing was removed from the property in the hopes that the *Eucalyptus spp.* overstorey would regenerate without active (and expensive) planting. Analyses of some related data have been published (Miles 2008, Vesk et al. 2016) but those analyses do not address the question analysts will answer in our study.

Researchers conducted three rounds of surveys at 18 sites across the Goulburn Broken catchment in northern Victoria, Australia in winter and spring 2006 and autumn 2007. In each survey period, a different set of 15 x 15 m quadrats were randomly allocated across each site within 60 m of existing tree canopies. The number of quadrats at each site depended on the size of the site, ranging from four at smaller sites to 11 at larger sites. The total number of quadrats surveyed across all sites and seasons was 351. The number of *Eucalytpus spp.* seedlings was recorded in each quadrat along with information on the GPS location, aspect, tree canopy cover, distance to tree canopy, and position in the landscape. Ground layer plant species composition was recorded in three 0.5 x 0.5 m sub-quadrats within each quadrat. Subjective cover estimates of each species as well as bare ground, litter, rock and moss/lichen/soil crusts were recorded. Subsequently, this was augmented with information about the precipitation and solar radiation at each GPS location.

Step 2: Recruitment and Initial Survey of Analysts

Initiating authors (TP, HF, SN, EG, SG, PV, FF) created a publicly available document providing a general description of the project (osf.io/mn5aj/). The project will be advertised at conferences, via Twitter, using mailing lists for ecological societies (full scope of these lists is not fixed but will include Ecolog, Evoldir, and lists for the Environmental Decisions Group and Transparency in Ecology and Evolution), and via word of mouth. The target population is active ecology, conservation, or evolutionary biology researchers with a graduate degree (or currently studying for a graduate degree) in a relevant discipline.

Researchers can choose to work independently or in a small team. For the sake of simplicity, we refer to these as 'analysis teams' though some may comprise one individual. Recruitment for this project is ongoing but we aim for a minimum of 12 analysis teams independently evaluating each dataset (see sample size justification below). We will simultaneously recruit volunteers to peer-review the analyses conducted by the other volunteers through the same channels. Our goal is to recruit a similar number of peer-reviewers and analysts, and to ask each peer reviewer to review a minimum of four analyses. If we are unable to recruit at least half the number of reviewers as analysis teams, we will ask analysts to serve also as reviewers (after they have completed their analyses). All analysts and reviewers will share co-authorship on this manuscript and will participate in the collaborative process of producing the final manuscript. All analysts will sign a consent (ethics) document (https://osf.io/xyp68/) approved by the Whitman College Institutional Review Board prior to being allowed to participate.

We identified our minimum number of analyses per data set by considering the number of effects needed in a meta-analysis to generate an estimate of heterogeneity (τ^2) with a 95% confidence interval that does not encompass zero. This minimum sample size is invariant regardless of τ^2. This is because the same *t*-statistic value will be obtained by the same sample size regardless of variance (τ^2). We see this by first examining the formula for the standard error, SE for variance, (τ^2) or SE(τ^2) assuming normality in an underlying distribution of effect sizes (Knight 2000):

$$SE(\tau^2) = \sqrt{\frac{2\tau^4}{(n-1)}}$$

and then rearranging the above formula to show how the t-statistic is independent of τ^2, as seen below.

$$t = \frac{\tau^2}{SE(\tau^2)} = \sqrt{\frac{(n-1)}{2}}$$

We then find a minimum *n* = 12 according to this formula.

Step 3: Primary Data Analyses

Analysis teams will register and answer a demographic and expertise survey (https://osf.io/seqzy/). We will then provide them with the dataset of their choice and request that they answer a specific research question. For the evolutionary biology dataset that question is "To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?" and for the ecology dataset it is "How does grass cover influence *Eucalyptus spp.* seedling recruitment?" Once their analysis is complete, they will answer a structured survey (https://osf.io/neyc7/), providing analysis technique, explanations of their analytical choices, quantitative results, and a statement describing their conclusions. They will also upload their analysis files (including the data set as they formatted if for analysis and their analysis code [if applicable]) and a detailed journal-ready statistical methods section.

Step 4: Peer Reviews of Analyses

At minimum, each analysis will be evaluated by four different reviewers, and each volunteer peer-reviewer will be randomly assigned methods sections from at least four analyst teams (the exact number will depend on the number of analysis teams and peer reviewers recruited). Each peer reviewer will register and answer a demographic and expertise survey identical to that asked of the analysts, except we will not ask about 'team name' since reviewers will not be working in teams. Reviewers will evaluate the methods of each of their assigned analyses one at a time in a sequence determined by the initiating authors (TP, HF, SN, EG, SG, PV, FF). The sequences will be systematically assigned so that, if possible, each analysis is allocated to each position in the sequence for at least one reviewer. For instance, if each reviewer is assigned four analyses to review, then each analysis will be the first analysis assigned to at least one reviewer, the second analysis assigned to another reviewer, the third analysis assigned to yet another reviewer, and the fourth analysis assigned to a fourth reviewer. Balancing the order in which reviewers see the analyses controls for order effects, e.g. a reviewer might be less critical of the first methods section they read than the last.

The process for a single reviewer will be as follows. First, the reviewer will receive a description of the methods of a single analysis. This will include the narrative methods section, the analysis team's answers to our survey questions regarding their methods, including analysis code, and the data set. The reviewer will then be asked, in an online survey (https://osf.io/4t36u/), to rate that analysis on a scale of 0-100 based on this prompt: "Rate the overall appropriateness of this analysis to answer the research question ([one of the two research questions inserted here]) with the available data. To help you calibrate your rating, please consider the following guidelines:

100. A perfect analysis with no conceivable improvements from the reviewer

75. An imperfect analysis but the needed changes are unlikely to dramatically alter outcomes

50. A flawed analysis likely to produce either an unreliable estimate of the relationship **or** an over-precise estimate of uncertainty

25. A flawed analysis likely to produce an unreliable estimate of the relationship **and** an over-precise estimate of uncertainty

0. A dangerously misleading analysis, certain to produce both an estimate that is wrong and a substantially over-precise estimate of uncertainty that places undue confidence in the incorrect estimate.

*Please note that these values are meant to calibrate your ratings. We welcome ratings of any number between 0 and 100."

After providing this rating, the reviewer will be presented with this prompt, in multiple-choice format: "Would the analytical methods presented produce an analysis that is (a) publishable as is, (b) publishable with minor revision, (c) publishable with major revision, (d) deeply flawed and unpublishable?" The reviewer will then be provided with a series of text boxes and the following prompts:

"Please explain your ratings of this analysis.

Please evaluate the choice of statistical analysis type.

Please evaluate the process of choosing variables for and structuring the statistical model.

Please evaluate the suitability of the variables included in (or excluded from) the statistical model.

Please evaluate the suitability of the structure of the statistical model.
Please evaluate choices to exclude or not exclude subsets of the data.
Please evaluate any choices to transform data (or, if there were no transformations, but you think there should have been, please discuss that choice)."
After submitting this review, a methods section from a second analysis will then be made available to the reviewer. This same sequence will be followed until all analyses allocated to a given reviewer have been provided and reviewed. After providing the final review, the reviewer will be simultaneously provided with all four (or more) methods sections that reviewer has just completed reviewing, the option to revise their original ratings, and a text box to provide an explanation. The invitation to revise the original ratings will be as follows: "If, now that you have seen all the analyses you are reviewing, you wish to revise your ratings of any of these analyses, you may do so now." The text box will be prefaced with this prompt: "Please explain your choice to revise (or not to revise) your ratings."

Step 5: Evaluate Variation

Initiating authors (TP, HF, SN, EG, SG, PV, FF) will conduct the analyses outlined in this section. We will describe the variation in model specification in several ways. We will calculate summary statistics describing variation among analysis, including mean, SD, and range of number of variables per model included as fixed effects, the number of interaction terms, the number of random effects, and the mean, SD, and range of sample sizes. We will also present the number of analyses in which each variable was included.

We will summarize the variability in standardized effect sizes and predicted values of dependent variables among the individual analyses using standard random effects meta-analytic techniques. First, we will derive standardized effect sizes from each individual analysis. We will do this for all linear models or generalized linear models by converting the *t* value and the degree of freedom (*df*) associated with regression coefficients (e.g. the effect of the number of siblings [predictor] on growth [response] or the effect of grass cover [predictor] on seedling recruitment [response]) to the correlation coefficient, *r*, using the following:

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

This formula will work only if *t* values and *df* are from linear models or generalized linear models (GLMs; Nakagawa and Cuthill 2007). If, instead, linear mixed-effects models (LMMs) and generalized linear mixed-effects models (GLMMs) were used by a given analysis, the exact *df* cannot be estimated. However, adjusted *df* can be estimated, for example, using the Satterthwaite approximation of *df,* or df_S (note that SAS uses this approximation to obtain df for LMMs and GLMMs; Luke 2017). For analyses using either LMMs or GLMMs that do not produce df_S, we will obtain df_S by rerunning the same (G)LMMs using the lmer or glmer function in the lmerTest package in R (Kuznetsova et al. 2017). Then, we will use *t* values and df_S from the models to obtain *r* as per the formula above. All *r* and accompanying *df* (df_S) will be converted to *Zr* and its sampling variance; $1/(n – 3)$ where $n = df + 1$. Any analyses from which we cannot derive a signed *Zr*, for instance one with a quadratic function in which the slope changes sign, will be excluded from the analyses of Fisher's *Zr*. We expect such analyses will be rare. Regardless, as we describe below, we will generate a second set of standardized effects (predicted values) that can be derived from any explanatory model produced by these data.

Besides *Zr*, which describes the strength of a relationship based on the amount of variation in a dependent variable explained by variation in an independent variable, we will also examine differences in the shape of the relationship between the independent and dependent variables. To accomplish this, we will derive a point estimate (predicted value) for the dependent variable of interest for each of three values of our primary independent variable (the 25th percentile, median, and 75th percentile), while setting the values of any co-variates at the corresponding quartiles (25th percentile, median, 75th percentile). We will use the 25th and 75th percentiles rather than minimum and maximum values to reduce the chance of occupying unrealistic parameter space. We will derive these predicted values from the model information provided by the individual analysts. All values (predictions) will be transformed to the original scale along with their standard errors (SE); the delta method (Ver Hoef 2012) will be used for the transformation of SE. We will use the square of the SE associated with predicted values as the sampling variance in the meta-analyses described below, and we will analyze these predicted values in exactly the same ways as we analyze *Zr* in the following analyses.

We will plot individual effect size estimates (*Zr*) and predicted values of the dependent variable (y) and their corresponding 95% confidence / credible intervals in forest plots to allow visualization of the range and precision of effect size and predicted values. Further, we will include these estimates in random effects meta-analyses (Higgins et al. 2003, Borenstein et al. 2017) using the metafor package in R (Viechtbauer 2010):

Zr ~ 1 + (1| analysis_identity)

and

$y_{median\ /\ 25th\ /\ 75th}$ ~ 1 + (1| analysis_identity)

where y is the predicted for the dependent variable at the 25th percentile, median, or 75th percentile of the independent variables. The individual *Zr* effect sizes will be weighted with the inverse of sampling variance for *Zr*. The individual predicted values for dependent variable (y) will be weighted by the associated SE^2. These analyses will provide an average *Zr* score or an average y with corresponding 95% confidence interval and will estimate two types of heterogeneity indices, τ^2 and I^2. The former, τ^2 is the absolute measure of heterogeneity or the between-study variance (in our case, between-team variance) whereas I^2 is a relative measure of heterogeneity. We obtain the estimate of relative heterogeneity (I^2) by dividing the between-analysis variance by the sum of between-analysis and within-analysis variance (sampling error variance). I^2 is thus the proportion of variance that is due to heterogeneity as opposed to sampling error. When calculating I^2, within-study variance is amalgamated across studies to create a "typical" within-study variance which serves as the sampling error variance (Higgins et al. 2003, Borenstein et al. 2017). Our goal here is to visualize and quantify the degree of variation among analyses in effect size estimates (Nakagawa and Cuthill 2007). We are not testing for statistical significance.

Finally, we will assess the extent to which deviations from the meta-analytic mean by individual effect sizes (*Zr*) or the predicted values of the dependent variable (y) are explained by the peer-rating of each analysis team's method section, by a measurement of the 'uniqueness' of the set of predictor variables included in each analysis, and possibly by the choice of whether or not to include random effects in the model. The deviation score, which serves as the dependent variable in these analyses, will be the absolute value of the difference between the meta-analytic mean *Zr* (or y) and the individual *Zr* (or y) estimate for each analysis. We will use the Box-Cox transformation on the absolute values of deviation scores to achieve an approximately normal distribution (c.f. Fanelli and Ioannidis 2013, Fanelli et al. 2017: supplement). We will describe variation in this dependent variable with both a series of univariate analyses and a multivariate analysis. All these analyses will be general linear (mixed) models. These analyses are secondary to our estimation of variation in effect sizes described above. We wish to quantify relationships among variables, but we have no a priori expectation of effect size and we will not make dichotomous decisions about statistical significance.

When examining the extent to which reviewer ratings (on a scale from 0 to 100) explain deviation from the average effect (or predicted value), each analysis will have been rated by multiple peer reviewers, so for each reviewer score to be included, we need to include each deviation score in the analysis multiple times. To account for the non-independence of multiple ratings of the same analysis, we will include analysis identity as a random effect in our generalized linear mixed model in the lme4 package in R (Bates et al. 2015). To account for potential differences among reviewers in their scoring of analyses, we will also include reviewer identity as a random effect:

Box-Cox(abs(Deviation score)) ~ rating + (1|analysis_identity) + (1|reviewer_identity),

where "Box-Cox" represents the Box-Cox and "abs" the absolute value. We will conduct a similar analysis with the four categories of reviewer ratings ((1) deeply flawed and unpublishable, (2) publishable with major revision, (3) publishable with minor revision, (4) publishable as is) set as ordinal predictors numbered as shown here. These analyses will also include random effects of analysis identity and reviewer identity. Both of these analyses (1: 1-100 ratings as the fixed effect, 2: categorical ratings as the fixed effects) will be conducted eight times for each data set, each of the four responses (*Zr*, y_{25th}, y_{median}, y_{75th}) will be compared once to the initial ratings provided by the peer reviewers, and again based on the revised ratings provided by the peer reviewers.

The next set of univariate analyses will seek to explain deviations from the mean effects based on a measure of the uniqueness of the set of variables included in each analysis. As a 'uniqueness' score, we will use Sorensen's Similarity Index (an index typically used to compare species composition across sites), treating variables as species and individual analyses as sites. To generate an individual Sorensen's value for each analysis requires calculating the pairwise Sorensen's value for all pairs of analyses (of the same data set), and then taking the average across these Sorensen's values for each analysis. We will calculate the Sorensen's index values using the betapart package (Baselga et al. 2018) in R:

$$\beta Sorensen = \frac{b+c}{2a+b+c}$$

where *a* is the number of variables common to both models, *b* is the number of variables that occur in the first model but not in the second and *c* is the number of variables that occur in the second model but not in the first. We then will use the per-model average Sorensen's index value as an independent variable to predict the deviation score in a general linear model, with no random effect since each analysis is included only once, in the stats package in R (R_Core_Team 2019):

Box-Cox(abs(Deviation score)) ~ βSorensen.

Next, we will assess the relationship between the inclusion of random effects in the analysis and the deviation from the mean effect size. We anticipate that most analysts will use random effects in a mixed model framework, but if we are wrong, we want to evaluate the differences in outcomes when using random effects versus not using random effects. Thus if there are at least 5 analyses that do and 5 analyses that do not include random effects, we will add a binary predictor variable "random effects included (yes/no)" to our set of univariate analyses and will add this predictor variable to our multivariate model described below.

Finally, we will conduct a multivariate analysis with the five predictors described above (peer ratings 0-100 and peer ratings of publishability 1-4; both original and revised and Sorensen's index, plus a sixth, presence /absence of random effects, if sample size is sufficient) with random effects of analysis identity and reviewer identity in the lme4 package in R. We will use the revised peer ratings only:

Box-Cox(abs(Deviation score)) ~ numerical_rating_1 + categorical_rating_1 + numerical_rating_2 + categorical_rating_2 + βSorensen + (1|analysis_identity) + (1|reviewer_identity).

We will conduct all the analyses described above eight times; for each of the four responses (*Zr*, y_{25th}, y_{median}, y_{75th}) one time for each of the two data sets.

We will publicly archive all relevant data, code, and materials on the Open Science Framework (osf.io). Archived data will include the original data sets distributed to all analysts, any edited versions of the data analyzed by individual groups, and the data we analyze with our meta-analyses, which include the effect sizes derive from separate analyses, the statistics describing variation in model structure among analyst groups, and the anonymized answers to our surveys of analysts and peer reviewers. Similarly, we will archive both the analysis code used for each individual analysis and the code from our meta-analyses. We will also archive copies of our survey instruments from analysts and peer reviewers.

Our rules for excluding data from our study are as follows. We will exclude from our synthesis any individual analysis submitted after we have completed peer review or those unaccompanied by analysis files that allow us to understand what the analysts did. We will also exclude any individual analysis that does not produce an outcome that can be interpreted as an answer to our primary question (as posed above) for the respective data set. For instance, this means that in the case of the data on blue tit chick growth, we would exclude any analysis that does not include something that can be interpreted as growth or size as a dependent (response) variable, and in the case of the Eucalyptus establishment data, we would exclude any analysis that does not include a measure of grass cover among the independent (predictor) variables. Also, as described above, any analysis that cannot produce an effect that can be converted to a signed *Zr* will be excluded from analyses of *Zr*.

Step 6: Facilitated Discussion and Collaborative Write-Up of Manuscript

Analysts and initiating authors will discuss the limitations, results, and implications of the study and collaborate on writing the final manuscript for review as a stage-2 Registered Report.

Ethics, consent and permissions

We have obtained permission to conduct this research from the Whitman College Institutional Review Board (IRB). As part of this permission, the IRB has approved the consent form (https://osf.io/xyp68/) that all participants will be asked to complete prior to joining the study.

The authors declare that they have no competing interests.

Literature Cited

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. 2015 **67**:48.

Borenstein, M., J. P. T. Higgins, L. Hedges, and H. Rothstein. 2017. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. Research Synthesis Methods **8**:5-18.

Botvinik-Nezer, R., F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, A. Adcock, P. Avesani, B. Baczkowski, A. Bajracharya, L. Bakst, S. Ball, M. Barilari, N. Bault, D. Beaton, J. Beitner, R. Benoit, R. Berkers, J. Bhanji, B. Biswal, S. Bobadilla-Suarez, T. Bortolini, K. Bottenhorn, A. Bowring, S. Braem, H. Brooks, E. Brudner, C. Calderon, J. Camilleri, J. Castrellon, L. Cecchetti, E. Cieslik, Z. Cole, O. Collignon, R. Cox, W. Cunningham, S. Czoschke, K. Dadi, C. Davis, A. De Luca, M. Delgado, L. Demetriou, J. Dennison, X. Di, E. Dickie, E. Dobryakova, C. Donnat, J. Dukart, N. W. Duncan, J. Durnez, A. Eed, S. Eickhoff, A. Erhart, L. Fontanesi, G. M. Fricke, A. Galvan, R. Gau, S. Genon, T. Glatard, E. Glerean, J. Goeman, S. Golowin, C. González-García, K. Gorgolewski, C. Grady, M. Green, J. Guassi Moreira, O. Guest, S. Hakimi, J. P. Hamilton, R. Hancock, G. Handjaras, B. Harry, C. Hawco, P. Herholz, G. Herman, S. Heunis, F. Hoffstaedter, J. Hogeveen, S. Holmes, C.-P. Hu, S. Huettel, M. Hughes, V. Iacovella, A. Iordan, P. Isager, A. I. Isik, A. Jahn, M. Johnson, T. Johnstone, M. Joseph, A. Juliano, J. Kable, M. Kassinopoulos, C. Koba, X.-Z. Kong, T. Koscik, N. E. Kucukboyaci, B. Kuhl, S. Kupek, A. Laird, C. Lamm, R. Langner, N. Lauharatanahirun, H. Lee, S. Lee, A. Leemans, A. Leo, E. Lesage, F. Li, M. Li, P. C. Lim, E. Lintz, S. Liphardt, A. Losecaat Vermeer, B. Love, M. Mack, N. Malpica, T. Marins, C. Maumet, K. McDonald, J. McGuire, H. Melero, A. Méndez Leal, B. Meyer, K. Meyer, P. Mihai, G. Mitsis, J. Moll, D. Nielson, G. Nilsonne, M. Notter, E. Olivetti, A. Onicas, P. Papale, K. Patil, J. E. Peelle, A. Pérez, D. Pischedda, J.-B. Poline, Y. Prystauka, S. Ray, P. Reuter-Lorenz, R. Reynolds, E. Ricciardi, J. Rieck, A. Rodriguez-Thompson, A. Romyn, T. Salo, G. Samanez-Larkin, E. Sanz-Morales, M. Schlichting, D. Schultz, Q. Shen, M. Sheridan, F. Shiguang, J. Silvers, K. Skagerlund, A. Smith, D. Smith, P. Sokol-Hessner, S. Steinkamp, S. Tashjian, B. Thirion, J. Thorp, G. Tinghög, L. Tisdall, S. Tompson, C. Toro-Serey, J. Torre, L. Tozzi, V. Truong, L. Turella, A. E. van't Veer, T. Verguts, J. Vettel, S. Vijayarajah, K. Vo, M. Wall, W. D. Weeda, S. Weis, D. White, D. Wisniewski, A. Xifra-Porxas, E. Yearling, S. Yoon, R. Yuan, K. Yuen, L. Zhang, X. Zhang, J. Zosky, T. E. Nichols, R. A. Poldrack, and T. Schonberg. 2019. Variability in the analysis of a single neuroimaging dataset by many teams. bioRxiv:843193.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretical approach. 2nd edition. Springer-Verlag, New York.

DeKogel, C. H. 1997. Long-term effects of brood size manipulation on morphological development and sex-specific mortality of offspring. Journal of Animal Ecology **66**:167-178.

Fanelli, D., R. Costas, and J. P. A. Ioannidis. 2017. Meta-assessment of bias in science. Proceedings of the National Academy of Sciences **114**:3714-3719.

Fanelli, D., and J. P. A. Ioannidis. 2013. US studies may overestimate effect sizes in softer research. Proceedings of the National Academy of Sciences **110**:15031-15036.

Fidler, F., Y. E. Chee, B. C. Wintle, M. A. Burgman, M. A. McCarthy, and A. Gordon. 2017. Metaresearch for evaluating reproducibility in ecology and evolution. BioScience **67**:282-289.

Forstmeier, W., E.-J. Wagenmakers, and T. H. Parker. 2017. Detecting and avoiding likely false-positive findings – a practical guide. Biological Reviews **92**:1941-1968.

Fraser, H., T. Parker, S. Nakagawa, A. Barnett, and F. Fidler. 2018. Questionable research practices in ecology and evolution. PLoS ONE **13**:e0200303.

Gelman, A., and D. Weakliem. 2009. Of beauty, sex, and power. American Scientist **97**:310-316.

Grueber, C. E., S. Nakagawa, R. J. Laws, and I. G. Jamieson. 2011. Multimodel inference in ecology and evolution: challenges and solutions. Journal of Evolutionary Biology **24**:699-711.

Higgins, J. P. T., S. G. Thompson, J. J. Deeks, and D. G. Altman. 2003. Measuring inconsistency in meta-analyses. BMJ **327**:557-560.

Klein, R. A., M. Vianello, F. Hasselman, B. G. Adams, R. B. Adams, S. Alper, M. Aveyard, J. R. Axt, M. T. Babalola, Š. Bahník, R. Batra, M. Berkics, M. J. Bernstein, D. R. Berry, O. Bialobrzeska, E. D. Binan, K. Bocian, M. J. Brandt, R. Busching, A. C. Rédei, H. Cai, F. Cambier, K. Cantarero, C. L. Carmichael, F. Ceric, J. Chandler, J.-H. Chang, A. Chatard, E. E. Chen, W. Cheong, D. C. Cicero, S. Coen, J. A. Coleman, B. Collisson, M. A. Conway, K. S. Corker, P. G. Curran, F. Cushman, Z. K. Dagona, I. Dalgar, A. Dalla Rosa, W. E. Davis, M. de Bruijn, L. De Schutter, T. Devos, M. de Vries, C. Doğulu, N. Dozo, K. N. Dukes, Y. Dunham, K. Durrheim, C. R. Ebersole, J. E. Edlund, A. Eller, A. S. English, C. Finck, N. Frankowska, M.-Á. Freyre, M. Friedman, E. M. Galliani, J. C. Gandi, T. Ghoshal, S. R. Giessner, T. Gill, T. Gnambs, Á. Gómez, R. González, J. Graham, J. E. Grahe, I. Grahek, E. G. T. Green, K. Hai, M. Haigh, E. L. Haines, M. P. Hall, M. E. Heffernan, J. A. Hicks, P. Houdek, J. R. Huntsinger, H. P. Huynh, H. IJzerman, Y. Inbar, Å. H. Innes-Ker, W. Jiménez-Leal, M.-S. John, J. A. Joy-Gaba, R. G. Kamiloğlu, H. B. Kappes, S. Karabati, H. Karick, V. N. Keller, A. Kende, N. Kervyn, G. Knežević, C. Kovacs, L. E. Krueger, G. Kurapov, J. Kurtz, D. Lakens, L. B. Lazarević, C. A. Levitan, N. A. Lewis, S. Lins, N. P. Lipsey, J. E. Losee, E. Maassen, A. T. Maitner, W. Malingumu, R. K. Mallett, S. A. Marotta, J. Međedović, F. Mena-Pacheco, T. L. Milfont, W. L. Morris, S. C. Murphy, A. Myachykov, N. Neave, K. Neijenhuijs, A. J. Nelson, F. Neto, A. Lee Nichols, A. Ocampo, S. L. O’Donnell, H. Oikawa, M. Oikawa, E. Ong, G. Orosz, M. Osowiecka, G. Packard, R. Pérez-Sánchez, B. Petrović, R. Pilati, B. Pinter, L. Podesta, G. Pogge, M. M. H. Pollmann, A. M. Rutchick, P. Saavedra, A. K. Saeri, E. Salomon, K. Schmidt, F. D. Schönbrodt, M. B. Sekerdej, D. Sirlopú, J. L. M. Skorinko, M. A. Smith, V. Smith-Castro, K. C. H. J. Smolders, A. Sobkow, W. Sowden, P. Spachtholz, M. Srivastava, T. G. Steiner, J. Stouten, C. N. H. Street, O. K. Sundfelt, S. Szeto, E. Szumowska, A. C. W. Tang, N. Tanzer, M. J. Tear, J. Theriault, M. Thomae, D. Torres, J. Traczyk, J. M. Tybur, A. Ujhelyi, R. C. M. van Aert, M. A. L. M. van Assen, M. van der Hulst, P. A. M. van Lange, A. E. van ’t Veer, A. Vásquez- Echeverría, L. Ann Vaughn, A. Vázquez, L. D. Vega, C. Verniers, M. Verschoor, I. P. J. Voermans, M. A. Vranka, C. Welch, A. L. Wichman, L. A. Williams, M. Wood, J. A. Woodzicka, M. K. Wronska, L. Young, J. M. Zelenski, Z. Zhijia, and B. A. Nosek. 2018. Many Labs 2: investigating variation in replicability across samples and settings. Advances in Methods and Practices in Psychological Science **1**:443-490.

Knight, K. 2000. Mathematical Statistics. Chapman and Hall, New York.

Kuznetsova, A., P. B. Brockhoff, and R. H. B. Christensen. 2017. lmerTest Package: tests in linear mixed effects models. 2017 **82**:26.

Luke, S. G. 2017. Evaluating significance in linear mixed-effects models in R. Behavior Research Methods **49**:1494-1502.

Miles, C. 2008. Testing market-based instruments for conservation in northern Victoria. Pages 133-146 *in* T. Norton, T. Lefroy, K. Bailey, and G. Unwin, editors. Biodiversity: Integrating Conservation and Production: Case Studies from Australian Farms, Forests and Fisheries. CSIRO Publishing, Melbourne, Australia.

Nakagawa, S., and I. C. Cuthill. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews **82**:591-605.

Nicolaus, M., S. P. M. Michler, R. Ubels, M. van der Velde, J. Komdeur, C. Both, and J. M. Tinbergen. 2009. Sex-specific effects of altered competition on nestling growth and survival: an experimental manipulation of brood size and sex ratio. Journal of Animal Ecology **78**:414-426.

Noble, D. W. A., M. Lagisz, R. E. O'Dea, and S. Nakagawa. 2017. Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses. Molecular Ecology **26**:2410-2425.

Open_Science_Collaboration. 2015. Estimating the reproducibility of psychological science. Science **349**:aac4716.

Parker, T. H., W. Forstmeier, J. Koricheva, F. Fidler, J. D. Hadfield, Y. E. Chee, C. D. Kelly, J. Gurevitch, and S. Nakagawa. 2016. Transparency in ecology and evolution: real problems, real solutions. Trends in Ecology & Evolution **31**:711-719.

R_Core_Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Royle, N. J., I. R. Hartley, I. P. F. Owens, and G. A. Parker. 1999. Sibling competition and the evolution of growth rates in birds. Proceedings of the Royal Society B-Biological Sciences **266**:923-932.

Senior, A. M., C. E. Grueber, T. Kamiya, M. Lagisz, K. O'Dwyer, E. S. A. Santos, and S. Nakagawa. 2016. Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. Ecology **97**:3293-3299.

Shavit, A., and A. M. Ellison, editors. 2017. Stepping in the same river twice: replication in biological research. Yale University Press, New Haven, Connecticut, USA.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. 2018. Many analysts, one data set: making transparent how variations in analytic choices affect results. Advances in Methods and Practices in Psychological Science **1**:337-356.

Simonsohn, U., J. P. Simmons, and L. D. Nelson. 2015. Specification curve: descriptive and inferential statistics on all reasonable specifications. Available at SSRN: https://ssrn.com/abstract=2694998.

Steegen, S., F. Tuerlinckx, A. Gelman, and W. Vanpaemel. 2016. Increasing transparency through a multiverse analysis. Perspectives on Psychological Science **11**:702-712.

Vander Werf, E. 1992. Lack's clutch size hypothesis: an examination of the evidence using meta-analysis. Ecology **73**:1699-1705.

Ver Hoef, J. M. 2012. Who invented the delta method? The American Statistician **66**:124-127.

Verhulst, S., M. J. Holveck, and K. Riebel. 2006. Long-term effects of manipulated natal brood size on metabolic rate in zebra finches. Biology Letters **2**:478-480.

Vesk, P. A., W. K. Morris, W. McCallum, R. Apted, and C. Miles. 2016. Processes of woodland eucalypt regeneration: lessons from the bush returns trial. Proceedings of the Royal Society of Victoria **128**:54-63.

Viechtbauer, W. 2010. Conducting meta-analyses in R with the metafor package. 2010 **36**:48.

Table 1. Timeline for completion of this project relative to timing of in-principle-acceptance (IPA) of stage 1 of the registered report.

Project step	Time to completion (after IPA)
Select data sets	completed
Recruitment and initial survey of analysts	Recruitment has begun. We will complete recruitment 2 months after IPA
Primary data analyses	4 months after IPA
Peer review of analyses	6 months after IPA
Evaluate variation	8 months after IPA
Facilitated discussion and collaborative write-up	11 months after IPA

Supplement

Table S1. The variable names and descriptions for all variables included in the blue tit data set.

Variable name	Variable description
chick_ring_number	Unique alpha numeric code on the aluminimum ring (band) fitted to the leg of the individual nestling. Each individual chick is listed in only one row in this datebase.
hatch_year	Year the nestling hatched
hatch_nest_breed_ID	Unique numeric code corresponding to the clutch of eggs from which this row's chick hatched. Each time a single clutch of eggs was laid in a given nest box, that clutch was assigned a breeding ID.
hatch_Area	Letter code corresponding to the area (a.k.a. 'compartment') of the study site in which the chick hatched
hatch_Box	Alpha numeric code corresponding to the individual nest box in which the chick was hatched. The code consists of the area code followed by the number of the nest box within that area.
hatch_mom_Ring	Unique alpha numeric code on the aluminimum ring (band) fitted to the leg of the adult female attending to the nest in which the chick hatched. This female can be assumed to have laid the eggs in this nest and thus to be the mother of this chick.
hatch_nest_dad_Ring	Unique alpha numeric code on the aluminimum ring (band) fitted to the leg of the adult male attending to the nest in which the chick hatched.
Extra-pair_paternity	Code indicating whether a molecular genetic paternity test identified the male attending the nest as the sire of the chick. "1" indicates a chick sired by a different ('extra-pair') male. "2" indicates a chick sired by the male attending the nest.
Extra-pair_dad_ring	For chicks identified as not sired by the adult male attending the nest, the identity (unique alpha numeric code on the aluminimum ring fitted to the leg) of the sire. Note that for many 'extra-pair' sired chicks, the identity of the sire was unknown (because the sire had not been captured and genotyped by us).
genetic_dad_ring_(WP_or_EP)	Ring number for the sire of the chick (whether the sire was extra-pair or within-pair)
hatch_nest_LD	The date of the first egg was laid ('lay date') in the clutch from which the chick hatched. 1 = April 1.
hatch_nest_CS	The size of the clutch (number of eggs) from which the chick hatched.
hatch_nest_OH	The date the eggs began to hatch in the clutch from which the chick hatched. 1 = April 1.

d0_hatch_nest_brood_size

The number of live chicks in the nest of hatching immediately after eggs have hatched (can be fewer than clutch size if one or more eggs did not hatch)

d14_hatch_nest_brood_size

In the nest where the chick hatched, the number of live chicks in the nest at day 14 after hatching

rear_nest_breed_ID

Unique numeric code corresponding to the brood of chick with which this row's chick was reared. Each time a single clutch of eggs were laid in a given nest box, that clutch was assigned a breeding ID.

rear_area

Letter code corresponding to the area (a.k.a. 'compartment') of the study site in which the chick was reared

rear_Box

Alpha numeric code corresponding to the individual nest box in which the chick was reared. The code consists of the area code followed by the number of the nest box within that area.

rear_mom_Ring

Unique alpha numeric code on the aluminimum ring (band) fitted to the leg of the adult female attending to the nest in which the chick was reared.

rear_dad_Ring

Unique alpha numeric code on the aluminimum ring (band) fitted to the leg of the adult male attending to the nest in which the chick was reared.

rear_nest_trt

Nest experimental manipulation treatment code for the nest in which the chick was reared. '5' indicates a net increase in the number of chicks (brood size increased), '6' indicates a net decrease in the number of chicks (brood size decreased), '7' indicates no manipulation (no chicks added or removed).

home_or_away

Indicates whether the chick was reared in the nest in which it hatched or had been moved by an experimenter to another nest. 1 = home (reared in nest of hatching); 2 = away (reared in transplanted nest)

rear_nest_LD

The date of the first egg was laid ('lay date') in the nest in which the chick was reared. 1 = April 1.

rear_nest_CS

The size of the clutch (number of eggs) in the nest in which the chick was reared

rear_nest_OH

The date the eggs began to hatch in the nest in which the chick was reared. 1 = April 1.

rear_d0_rear_nest_brood_size

The number of live chicks in the nest of rearing immediately after eggs have hatched (before chicks were moved among nests for the experiment)

rear_Cs_out

Number of chicks removed by experimenters from nest in which the chick was reared

rear_Cs_in

Number of chicks added by experimenters to nest in which the chick was reared

net_rearing_manipulation

Net change in chick number in nest where chick reared

rear_Cs_at_start_of_rearing	Number of chicks in nest of rearing immediately following experimental chick removals and additions
d14_rear_nest_brood_size	In the nest where the chick was reared, the number of live chicks in the nest at day 14 after hatching
number_chicks_fledged_from_rear_nest	In the nest where the chick was reared, the number of chicks that survived to leave the nest
Date_of_day14	date (1 = April 1) of 14th day after hatching
day_14_tarsus_length	Length of chick tarsometatarus in mm at day 14 after hatching
day_14_weight	mass of chick in grams at day 14 after hatching
day14_measurer	Code corresponding to the identity of the person who measured the chick at day 14.
chick_sex_molec	Sex of the chick based on molecular genetic analysis. 1 = female; 2 = male. Not all chicks were sexed.
chick_survival_to_first_breed_season	Indicates whether the chick was documented as attempting to breed on the study site in any subsequent year

Table S2. The variable names and descriptions for all variables included in the Eucalyptus recruitment data set.

Variable name	Variable description
SurveyID	Unique ID for each survey
Date	Date: DD/MM/YYYY
Season	Season
Property	The site that the surveys are associated with (identified by land owner's surname)
Quadrat_no	At each property/site in each sampling season the quadrats are numbered 1-n where n is the total number of quadrats/transects/sampling units undertaken at a site. N is dictated by the amount of relevant vegetation at each property.
Easting	Easting coordinate
Northing	Northing coordinate
Aspect	Aspect at each quadrat. Is it facing north (n), south (s), east (e), west, north east (ne) etc
Landscape_position	Is the quadrat located on a slope, on the flat, at the crest of a hill, at the toe of the slope (where the slope levels out)
ExoticAnnualGrass_cover	Percentage cover of exotic annual grass per quadrat
ExoticAnnualHerb_cover	Percentage cover of exotic annual herbacious plants per quadrat
ExoticPerennialHerb_cover	Percentage cover of exotic perennial herbacious plants per quadrat
ExoticPerennialGrass_cover	Percentage cover of exotic perennial grass per quadrat
ExoticShrub_cover	Percentage cover of woody plants less than 2m tall per quadrat
NativePerennialFern_cover	Percentage cover of native perennial ferns per quadrat
NativePerennialGrass_cover	Percentage cover of native perennial grasses per quadrat
NativePerennialHerb_cover	Percentage cover of native perennial herbs per quadrat
NativePerennialGraminoid_cover	Percentage cover of native perennial grasslike species that are not technically grasses (e.g. lilies) per quadrat
NativeShrub_cover	Percentage cover of native woodland plants less than 2m tall per quadrat
BareGround_cover	Percentage cover of bare ground (ground not covered by plants of leaf litter) per quadrat
Litter_cover	Percentage cover of leaf litter per quadrat
MossLichen_cover	Percentage cover of moss and lichen per quadrat
Rock_cover	Percentage cover of rock (as a substrate, not including small, moveable pebbles on soil etc) per quadrat
Euc_canopy_cover	Percentage foliage projective cover by eucalypt canopy per quadrat

Distance_to_Eucalypt_canopy(m)	Distance between the quadrat and the closest edge of the nearest eucalypt canopy in meters
euc_sdlgs0_50cm	The number of eucalypt seedlings between 0 and 50cm tall per quadrat
euc_sdlgs50cm-2m	The number of eucalypt seedlings between 50cm and 2m tall per quadrat
euc_sdlgs>2m	The number of eucalypt seedlings greater than 2m tall per quadrat
annual_precipitation	Total annual precipitation at the quadrat easting and northing in the sample year. Derived from Bureau of Meteorology data
precipitation_warmest_quarter	Total precipitation in the warmest quarter of the year at the quadrat easting and northing in the sample year. Derived from Bureau of Meteorology data
precipitation_coldest_quarter	Total precipitation in the coldest quarter of the year at the quadrat easting and northing in the sample year. Derived from Bureau of Meteorology data
PET	Potential evapotranspiration at each quadrat easting and northing.
MrVBF	Multi-resolution Valley Bottom Flatness index at the quadrat easting and northing : index of valley bottom flatness at scales from 10m to 10^5m to distinguish hillslopes from valley bottoms. Derived from a gridded MrVBF surface computed from a 20m DEM using an algorithm described in (Gallant & Dowling 2003)
K_perc	Estimated potassium (K) concentrations (%) in the top 30–45cm of the earth's crust at quadrat easting and northing. Derived from ‘GSV Wangaratta South Vic magnetic line data’ (1997) Geophysical Archive Data Delivery System, Department of Primary Industries, Victoria. Estimated from airborne Gamma radiometric spectrometry surveys
Th_ppm	Estimated thorium (Th) concentrations (ppm) in the top 30–45cm of the earth's crust at quadrat easting and northing. Derived from ‘GSV Wangaratta South Vic magnetic line data’ (1997) Geophysical Archive Data Delivery System, Department of Primary Industries, Victoria. Estimated from airborne Gamma radiometric spectrometry surveys
U_ppm	Estimated uranium (U) concentrations (ppm) in the top 30–45cm of the earth's crust at quadrat easting and northing. Derived from ‘GSV Wangaratta South Vic magnetic line data’ (1997) Geophysical Archive Data Delivery System, Department of Primary Industries, Victoria. Estimated from airborne Gamma radiometric spectrometry surveys
SRad_Jan	Incoming solar radiation (WH/m2) to quadrat easting and northing in January of the sample year. Computed using the solar radiation analysis tools in the Spatial Analyst extension for ArcGIS 9.2 Desktop.
SRad_Jul	Incoming solar radiation (WH/m2) to quadrat easting and northing in July of the sample year. Computed using the solar radiation analysis tools in the Spatial Analyst extension for ArcGIS 9.2 Desktop.