Group Processes & Intergroup Relations

G P I R

Article

Group Processes & Intergroup Relations
2015, Vol. 18(3) 329–347
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1368430214558311
gpir.sagepub.com
SAGE

Maximum-likelihood methods for meta-analysis: A tutorial using *R*

Jack L. Vevea[1] and Kathleen M. Coburn[1]

Abstract

The method of maximum likelihood provides a versatile way to estimate and conduct inference about moderators of effect size in meta-analytic models. The *metafor* package for the open-source statistical software *R* offers easy access to this method. We discuss inferential choices that the meta-analyst must make, and advocate the general choice of random-effects methods. We demonstrate the use of the *metafor* package using data from two meta-analyses that address group processes. These demonstrations illustrate two contrasting approaches to meta-analytic inference: a priori random-effects inference and conditionally random inference. The examples show that these approaches typically lead to the same conclusions when applied correctly.

Keywords

maximum likelihood, meta-analysis, mixed effects, open source, random effects

Paper received 16 April 2014; revised version accepted 15 September 2014.

Meta-analysis has become well established as a method for integrating what is known about a field. Researchers in a wide variety of areas employ the technique. Prominent fields that use meta-analysis include medicine (e.g., integrating clinical trials; DerSimonian & Laird, 1986), ecology and evolutionary biology (e.g., Koricheva, Gurevitch, & Mengersen, 2013), criminology (e.g., Wilson, 2001), education, which provided the context for the coinage of the term "meta-analysis" (e.g., Glass, 1976; Hattie, Rogers, & Swaminathan, 2014), and psychology, where the approach first generated wide attention (Smith, Glass, & Miller, 1980). Meta-analytic methods have evolved from an early emphasis on calculating a simple estimate of a true effect to a richer set of techniques that allow the researcher to address questions about factors that are systematically related to variation in effect magnitude. Such factors may be characteristics of how research was operationalized (e.g., variation in the strength of association in a prisoner's dilemma reward matrix; Wildschut, Pinter, Vevea, Insko, & Schopler, 2003), characteristics of the populations studied (e.g., do the effects of violent video games differ for Western and Japanese populations?; Anderson et al., 2010), characteristics of researchers (e.g., gender of authors), characteristics

[1] The University of California, Merced, USA

Corresponding author:
Jack L. Vevea, Psychological Sciences, School of Social Sciences, Humanities and Arts, The University of California, Merced, 1100 North Lake Road, Merced, CA 95343, USA.
Email: jvevea@ucmerced.edu

of the primary studies being synthesized (e.g., published or not, funded or not, effect is of primary or secondary interest), and so on.

As the questions addressed by meta-analysis have grown more complex, common practice for analysis has tended to lag behind best practice. Although methodological guides are available that describe current analytic techniques (e.g., Cooper, Hedges, & Valentine, 2009), many researchers continue to perform meta-analyses by disaggregating effect sizes into separate groups depending on values of moderators and conducting separate analyses for those groups. This approach has distinct disadvantages compared to methods that characterize effect-size variation in a single model, including study characteristics and other factors as predictors.

The method of maximum likelihood is a general tool that provides a straightforward means of estimating the parameters of complex models. The idea is to select parameter estimates that maximize the likelihood of the data given the parameters. (The likelihood is simply the joint density viewed as a function of the parameters rather than of the data.) All researchers are familiar with maximum likelihood estimates (MLEs), whether or not they know they are. For example, the sample mean is a maximum likelihood estimate of the population mean if the population is normally distributed. Ordinary least squares regression estimates are MLEs if the distribution of errors is normal. The biased variance estimate (the sum of squared deviations from the mean divided by N instead of by $N - 1$) is a maximum likelihood estimate. This last example demonstrates a common characteristic of MLEs: they are often biased. However, the sampling distributions of MLEs can be shown to be less variable than those of any alternative estimates, and the reduced sampling variability is often sufficient to compensate for any bias, so that the estimates are closer on average to the true parameter values.

The purpose of this paper is to provide an introduction to the use of maximum likelihood estimation for meta-analysis. We assume a basic familiarity with the principles of meta-analysis; in particular, we do not discuss the processes of searching the literature, coding, and extracting effect sizes. We demonstrate maximum likelihood estimation using a package known as *metafor* (Viechtbauer, 2010), which is available for the open-source statistical software *R* (R Core Team, 2013). We chose *R* partly because the *metafor* package does a particularly good job of implementing maximum likelihood estimation for meta-analysis, but also because *R* is freely available to all readers. Thus we avoid presenting extended parallel examples in *SPSS, SAS, Stata*, and so on. After some preliminary treatment of *R* basics, we discuss the issue of selection of fixed- or random-effects models for meta-analysis. Finally, we demonstrate various analytic techniques using two examples that are relevant to group processes research. The first of these (Li Lu, Yuan, & McLeod, 2012) investigates hidden profiles in group decision making in a comparatively small meta-analysis (33 effect sizes). The second (Pettigrew & Tropp, 2006) deals with intergroup contact theory, and the meta-analysis includes 713 effect sizes. In the current paper, we work with an arbitrary sample of slightly more than half of those effects.

R Basics

Obtaining R

In order to implement the techniques presented in this paper, the reader first needs to obtain the *R* software. *R* is available at a website known as *CRAN* (which stands for *Comprehensive R Archive Network*; http://cran.us.r-project.org/). At the top of that page are links to install *R* under Linux, Mac OS X, and Windows. Installing the "base" *R* is a seamless process, similar to installing any other software.

Getting Data Into R

Once *R* is installed on your system, you may start it by double clicking the icon for the software. (This may be in various places depending on your operating system and on choices you made during installation.) You will see a console window and a menu at the top of the screen. The ">" in the console window is a command prompt indicating that *R* is waiting for you to enter an instruction. In any statistical software, the most basic requirement is to be able to enter data. There are various ways to accomplish this in *R*. The simplest is to use the "<-" assignment operator. For example, entering "MyVariable <- c(1,3,5,7,9)"[1] will assign the odd numbers from 1 to 9 to a variable called "MyVariable." (Note that variable names are case sensitive.) Once a variable has values, one can perform various operations on it using simple, usually intuitive commands. For example, if we type "MyVariable" at the command prompt, *R* will print the values 1, 3, 5, 7, and 9. If we type "mean(Myvariable)", *R* will print the value 5, which is the mean of those numbers. Similarly, "sd(MyVariable)", "median(MyVariable)" and "var(MyVariable)" return the sample standard deviation, median, and variance.

Entering data for a meta-analysis will usually involve a number of variables, and it will be most practical to read the data from a file. *R* has multiple ways of accomplishing this, but we will focus on the "read.table()" command. The syntax is "dataname <-read.table("*path/to/file/mydata.txt*")" where "dataname" is the desired name of the dataset and "mydata.txt" is a space- or tab-delimited text file. For example, if we were to save the contents of Table 1 on the desktop of a Windows machine in a text file named "LiLu.txt," the command "LiLu <- read.table("c:/users/username/desktop/LiLu.txt")" would read the data and save them in a variable named "LiLu." Complex, multicomponent variables such as this are referred to as "data frames" in *R*. If we look at LiLu by typing "LiLu" at the command prompt, we see that *R* has arbitrarily named the columns V1, V2, etc. To separate them, one can refer, for example, to "LiLu$V1". It will be more useful if the columns of the table are informatively named so that we can refer intuitively to individual components of the data frame. One way to accomplish this is to add column names to the read.table()command: "LiLu <- read.table("c:/users/username/desktop/LiLu.txt", col.names=c("d","v","groupsize", "infoload","task"))". Now we can refer to the effect size estimates as "LiLu$d", or to the sampling variances of the effect sizes as "LiLu$v". Better yet, we can issue the command "attach(LiLu)" and "d" or "v" will refer directly to the components of LiLu. Another option for specifying column names is to include the name at the head of each column in the text file, and add "header=TRUE" to the read.table() command, like this: read.table("c:/users/username/desktop/LiLu.txt",header=TRUE)".

Loading Packages in R

One of the remarkable features of *R* is that it is infinitely extendable. Users can easily write their own functions, and there is a large community of persons who create extensions for specific purposes. Added functions are often bundled into packages, which other users can load and use. The software we will use for meta-analysis, *metafor*, is one such package. In order to use packages, it is first necessary to install and load them. The installation is easily accomplished through a sequence of menu choices; users need only navigate to the "packages" drop-down menu, then select "install packages." Once the user clicks "install packages," a window for selecting a mirror site will appear. Choose one that is geographically nearby. Finally, after specifying a mirror site, the user sees a list of available packages and selects the desired one (in this case, *metafor*).

Once the user has installed a package, it is permanently available for use in *R*. However, when one has installed the package for the first time, and subsequently each time one begins a new *R* session, it is necessary to load the package. For *metafor*, the following command, issued at the *R* command prompt, will accomplish this: "library(metafor)". Remember that you will need to issue this library command each time you have started *R* before you can use the package.

Some Conceptual Basics for Meta-Analysis

Although many meta-analyses have proceeded by disaggregating effects into separate groups according to the value of potential moderating variables, a more flexible approach is to build a model, much as a primary researcher might employ ANOVA, regression, or ANCOVA models. There are a

Table 1. Data taken from Li Lu et al. (2012).

Standardized Mean difference	Sampling variance	Group size	Information load	Task
0.34	0.081	3	24	1
0.94	0.135	3	24	1
0.65	0.128	3	24	1
3.84	0.474	3	31	0
2.14	0.123	−99	36	0
1.87	0.411	10	76	0
2.27	0.164	6	54	0
3.16	0.150	4	700	1
2.35	0.121	4	30	0
1.09	0.036	3	30	0
8.00	0.947	3	45	0
1.25	0.129	3	29	0
2.16	0.135	4	−99	0
0.71	0.095	3	60	0
0.61	0.072	3	50	0
2.93	0.276	3	40	0
1.70	0.143	3	−99	0
3.70	0.417	3	−99	0
2.77	0.121	3	18	1
1.63	0.161	3	9	0
4.11	0.254	4	50	0
5.13	0.286	3	45	0
4.56	0.240	3	45	0
1.29	0.046	3	45	0
2.09	0.132	3	60	0
2.17	0.138	4	42	0
1.21	0.089	3	36	0
1.66	0.048	3	40	0
2.44	0.233	3	28	1
2.34	0.241	3	28	1
3.67	0.275	3	35	0
0.54	0.086	3	36	1
0.66	0.062	3	24	0

number of advantages to taking a model-based approach. It is possible to include multiple moderators that might work together to explain variation in effect magnitude. It is possible to consider interactions between moderators. It is possible to include moderators for purposes of statistical control (e.g., adding socioeconomic status of the study sample). One may mix categorical and continuous moderators. All of these models are very much analogous to the common linear techniques with which we are all familiar from primary research.

What differs for meta-analysis is the error structure of the models. Meta-analysis can be performed on a variety of effect-size measures: correlations, transformed correlations, standardized differences between means, and log odds ratios are most commonly employed in psychology. The techniques we demonstrate here may be used with any of these effect measures.[2] Formulas for the sampling variances of various measures are given in any introductory meta-analysis text. For each measure, sampling variance is primarily or entirely determined by sample size. For that reason, sampling variances are treated as essentially known, rather than estimated as in conventional statistical analysis. Hence, if one is interested in an ANOVA-like model for effects, a conventional ANOVA is inappropriate, as it treats the known variance as something to be estimated. Instead, meta-analytic models account for this special knowledge about sampling variability.

This known error structure simplifies inference. The *t* statistic of primary research becomes a *z* statistic. Hypotheses that would ordinarily be tested by *F* statistics now are tested with chi-square statistics. Moreover, the knowledge about sampling variances of effect sizes makes it simple to assess whether observed estimates are more variable than would be expected if sampling variability were the only source of variation. If effect sizes are more variable than expected, the excess variation may be quantified as a between-studies variance component.

Selecting an Inference Model for Meta-Analysis

Addressing excess variation raises the question of the choice of an inference model for meta-analysis. The basic choice is between fixed-effects and random-effects inference. Fixed-effects inference models the uncertainty associated with the sampling of data units (for psychology, typically persons) into studies, and conducts inferences relevant to the question of what would happen if new units were sampled into exactly the same set of studies. In contrast, random-effects inference considers that source of uncertainty, but also treats the studies themselves as samples from a larger universe of possible studies, and estimates an additional component of uncertainty (called a *variance component*) associated with the sampling of studies. The goal of inference in random-effects models is to generalize to what would happen in future studies.

In addition to affecting inference, the choice of a random-effects model effects our conceptual approach to estimation. In the fixed-effects model, we think about estimating the true mean effect (or, in more complex analyses, the true mean effect conditional on moderating variables). In the random-effects model, we think of estimating the mean and variance of a *distribution* of true mean effects (or, in more complex models, the mean of a conditional distribution given the values of moderators). The additional uncertainty associated with estimating the variance of that distribution reduces the power of random-effects inference and diminishes the weight given to the larger effects (simultaneously increasing the weight of the smallest effects).

The choice of inference model is not without controversy. For example, Bonett (2009) argues that because the effect sizes available to the meta-analyst are not a random sample from a population of effect sizes, random-effects models are inappropriate. Our own position is that exactly the same criticism could be levied against most primary research endeavors: few of us ever work with true random samples of persons, yet we are willing to act as though our results generalize to some population of interest. To the degree that consensus is possible in a field as broad as meta-analysis, there appears to be growing agreement among most practitioners applying the methods in a variety of fields that random-effects models are what is needed to support conclusions that are relevant to the world at large. For more detailed discussion of these issues, see Hedges and Vevea (1998) and Schmidt, Oh, and Hayes (2009).

We unabashedly take the stand that random-effects models are almost always appropriate for meta-analysis. In practice, though, there will usually be moderators of effect magnitude that warrant consideration. A fundamental question in meta-analysis is how to approach inference about these moderators. One possible choice, which at one time was the only solution frequently observed in practice, is to explore the influence of moderators in the context of a fixed-effects model. Another approach, which we believe to be more sound, is to add fixed-effects moderators to a random-effects model, resulting in a model that is often referred to as "mixed-effects." A third approach is to follow a sequence of steps that makes the selection of inference model an inferential choice in its own right, also known as a "conditionally random" approach. This method begins with a simple fixed-effects estimate of the mean effect. The meta-analyst then performs a heterogeneity test (typically a Q statistic that has a chi-square distribution if the only source of effect-size variation is the sampling of persons into studies). If that heterogeneity test is significant, the analyst tests moderators that might explain heterogeneity in effect sizes. If significant residual heterogeneity still remains in the presence of the moderators, then the analyst includes a variance component, resulting in a mixed-effects model. Often when that variance component is added, moderators that were significant according to fixed-effects tests are no longer significant. Most introductory texts on how to conduct a meta-analysis advocate some form of this conditional approach; see, for example, Becker and Schram (1994), Cooper (2010), Hartung, Knapp, and Sinha (2008), Hedges (1994), Hedges and Olkin (1985), Konstantopoulos and Hedges (2009), Leandro (2007), Shadish and Haddock (1994, 2009), Sutton, Abrams, Jones, Sheldon, and Song (2000), and Wang and Bushman (1999).

In each of our following examples, we will illustrate both the a priori mixed-effects approach and the conditionally random approach. The two methods usually result in the same final model; however, that sometimes requires the analyst to reverse the decision about the effect of a moderator that was significant in the fixed-effects context but is not necessarily so in the mixed-effects context.

Examples

We illustrate the conduct of meta-analysis using two examples from the group processes literature. In each case, we use only part of the original meta-analysis. For the smaller dataset, we focus on one question of several that were presented in the original paper, and we work with only a subset of the moderator variables that were employed. For the larger dataset, we also select only a subset of the original moderating variables, and we arbitrarily truncate the data. Our purpose in these changes is to focus on the process of conducting a meta-analysis and not on the substantive research questions embodied in the two papers. In particular, it is not our intent to reconsider or call into question any of the conclusions of the original meta-analyses.

Li Lu et al. (2012)

Our first example employs data from a meta-analysis conducted by Li Lu et al. (2012) that examines the effects of the hidden profile paradigm on group decision making. The hidden profile paradigm is a method of information distribution in which some information is shared by all group members prior to group discussion and some "unique" information is known only to select members. The common information will result in a suboptimal decision path, so it is only by collaborating and revealing the unique information that groups can reach an optimal decision. Ideally, groups should mention more unique information than common information. Certain theories, like the persuasive arguments theory of influence (Burnstein & Vinokur, 1977), posit that because novel arguments are more influential, group members will be more interested in hearing unique information, and will readily discover the "hidden profile." However, many studies find that groups fail at uncovering the unique information, and spend much more time discussing the common information (Li Lu et al., 2012). The authors summarize the findings of 65 studies that use this hidden profile paradigm.

The framework for understanding information sharing under the hidden profile paradigm has several components, which the meta-analysis examines separately. We focus here on a dataset of standardized mean differences (d) that represent the difference between common and unique information mentioned during discussion. The authors assess the impact of five moderator variables on unique information sharing: group size, percentage of unique information available, total information load, task type, and hidden profile strength. Here, for simplicity, we confine our interest to group size, information load, and task type.

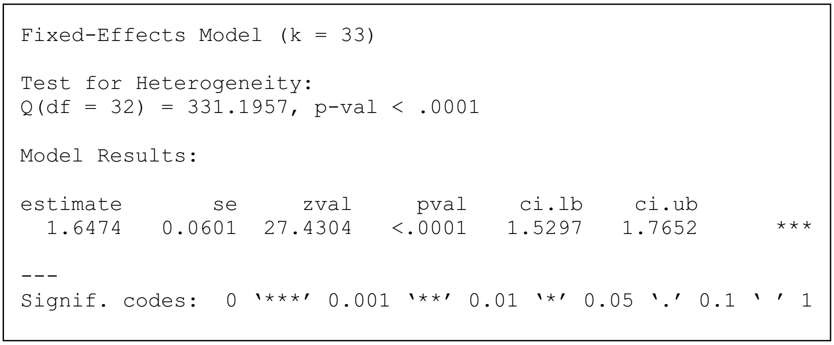
Past research on the impact of group size has been inconclusive (Li Lu et al., 2012). Increasing group size might lead to an increase in sharing of unique information because each group member will have a reduced cognitive load; however, it is also possible that more common information will be discussed, because the increased number of people means more people can bring up the same piece of information, and because larger groups increase the likelihood of social loafing. Amount of total information load is related to group members' cognitive load such that group members with reduced cognitive load should be more likely to share unique information. Task type indicates whether the solution has high demonstrability (coded as 1) or low demonstrability (coded as 0); solutions in which the full set of information leads to an unequivocal correct answer (e.g., math problems) have high demonstrability, while solutions that still vary even with the full set of information (e.g., voting decisions) have low demonstrability. Groups are theorized to share more unique information when solutions have high demonstrability.

Results presented in the current paper will not necessarily match those of Li Lu et al. (2012). This is partly because we use different methods and software, but also because the 2012 paper provides only a single sample size for each study with no information about the *n* in each group. We calculate sampling variances by assuming the *n* in each group is exactly half of the reported *N*, resulting in variances that are not identical to those used in the original analyses.

Preliminary analysis. Table 1 presents the data extracted from the paper by Li Lu et al. (2012).[3] We begin our analysis by reading the data into *R*, as explained before: "LiLu <- read.table("c:/users/username/desktop/LiLu.txt", col.names=c("d"," v","groupsize","infoload","task"))" and attach the dataset for easy reference ("attach(LiLu)"). Next, we load the *metafor* package: "library(metafor)". The primary tool we will use for conducting meta-analysis is the function "rma."

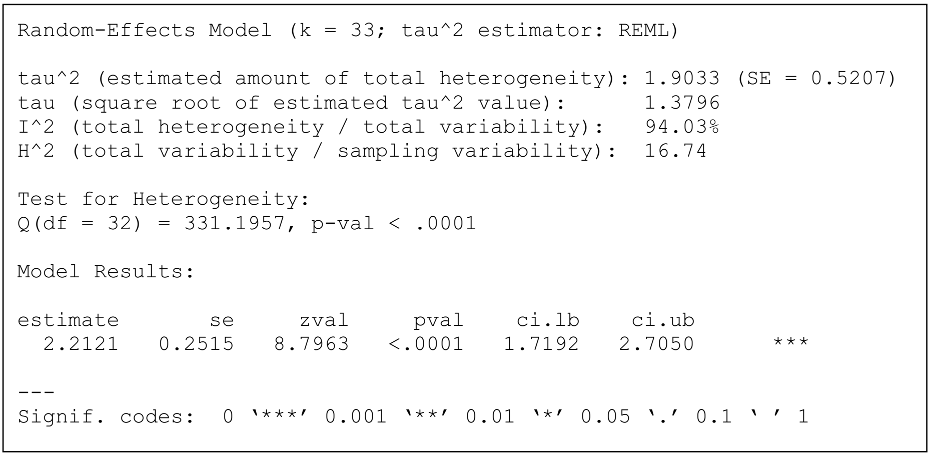
Although we argue that an a priori choice of random-effects models is usually appropriate, we start with a fixed-effects analysis because we intend to demonstrate both mixed-effects inference and conditionally random inference, and a fixed-effects model is the starting point for conditionally random inference. The command "rma(d,v,method='FE')" will conduct a simple fixed-effects meta-analysis. Figure 1 shows the results. We see that although the estimated mean effect is large (1.65) and highly significant (z = 27.43, $p < .0001$), the dataset is heterogeneous. The *Q-within* statistic that assesses whether effect sizes are more variable than would be expected if the only source of variation were sampling variability is highly significant ($QE = 331.20$, $df = 32$, $p < .0001$). Next, we estimate a simple random-effects analysis ("rma(d,v)"). The results appear in Figure 2. The estimated mean effect has increased in size (2.21), remains significant (z = 8.35, $p < .0001$), and differs in magnitude from the fixed-effects mean due to the inclusion of a variance component. The presence of a variance component tends to reduce the degree to which larger studies are weighted in the analysis, so that effect estimates can change substantially, as has occurred in this instance. Notice that the same *Q* test for heterogeneity appears; now, however, it represents a test of the null hypothesis that the variance component is zero. The variance component ("tau^2" in the output) is estimated to be 1.90. The I^2 value of 94.03 percent indicates that about 94% of the total observed variability in effect sizes is associated with variation in the population of random effects rather than sampling variation.

Mixed-effects inference. The meta-analyst who desires to generalize his or her findings to the universe of possible studies of the issue would select a random-effects approach. Given that we have possible moderators of effect size, these would be added in a mixed-effects model (or a series of such models). Our simplest potential moderator is task type, which is a binary variable coded 0 or 1 depending on whether the task has low or high demonstrability. To estimate the effect of



```
Fixed-Effects Model (k = 33)

Test for Heterogeneity:
Q(df = 32) = 331.1957, p-val < .0001

Model Results:

estimate       se      zval      pval    ci.lb    ci.ub
  1.6474   0.0601   27.4304   <.0001   1.5297   1.7652      ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1. A simple fixed-effects analysis of a heterogeneous dataset.

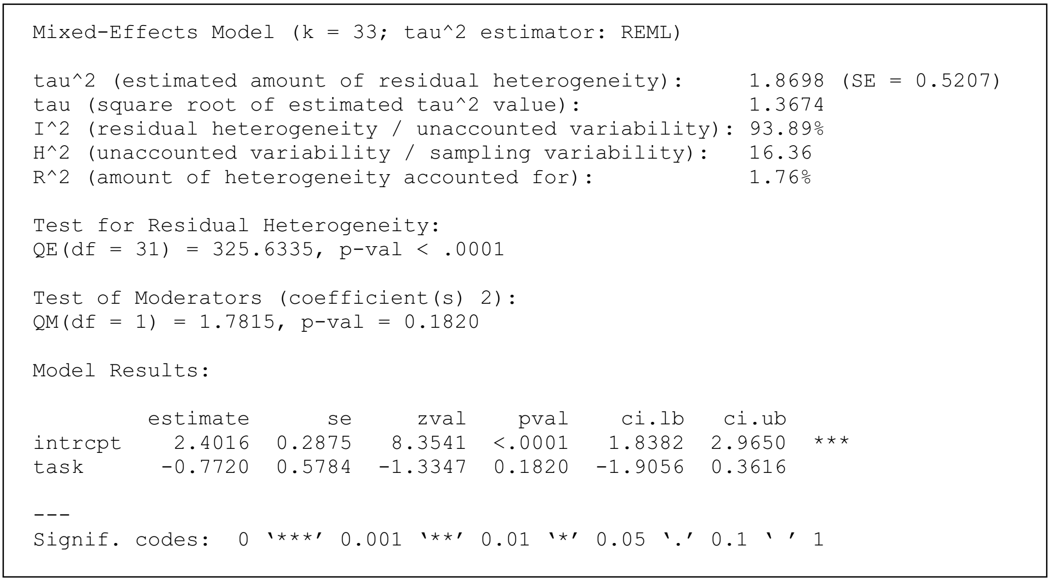


```
Random-Effects Model (k = 33; tau^2 estimator: REML)

tau^2 (estimated amount of total heterogeneity): 1.9033 (SE = 0.5207)
tau (square root of estimated tau^2 value):      1.3796
I^2 (total heterogeneity / total variability):   94.03%
H^2 (total variability / sampling variability):  16.74

Test for Heterogeneity:
Q(df = 32) = 331.1957, p-val < .0001

Model Results:

estimate       se      zval      pval    ci.lb    ci.ub
  2.2121   0.2515    8.7963   <.0001   1.7192   2.7050      ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2. A simple random-effects analysis of a heterogeneous dataset.

this moderator, we fit a linear model of the form $d = \beta_0 + \beta_1 I_{high}$, where I_{high} is the binary variable. To accomplish this in *metafor*, we enter "rma(d~task,v)"; the results appear in Figure 3. The estimated intercept for this model (2.4016) represents the mean effect for the low-demonstrability group, which was coded as 0 for the purposes of this analysis. The estimated slope (−0.772) is the difference in the mean effect for the high-demonstrability group; hence, the mean effect for that group is 2.4016 − 0.772 = 1.63. That smaller effect is consistent with theory; however, the slope is not significantly different from zero ($z = -1.33$, $p = .18$). Notice that the estimated variance component is slightly smaller than in the simple random-effects model (1.87 vs. 1.90), and that the heterogeneity is still significant ($QE = 325.63$, $df = 31$, $p < .0001$). The reported *QM* statistic is redundant here; in cases like this, with one degree of freedom, it is equivalent to the square of the z test for the slope.

Next, consider the continuous predictor, size of group. A problem arises because the fifth case is coded −99, indicating missing data. Although sophisticated techniques for handling missing data are available for meta-analysis (see, for example, Pigott, 2009), they are beyond the scope of this paper; we will address missing data by listwise deletion. *R* provides a function to divide variables according to the results of a logical test. Here, we will need to use this to delete the fifth case for every variable in our model. We can accomplish this by using the



```
Mixed-Effects Model (k = 33; tau^2 estimator: REML)

tau^2 (estimated amount of residual heterogeneity):     1.8698 (SE = 0.5207)
tau (square root of estimated tau^2 value):             1.3674
I^2 (residual heterogeneity / unaccounted variability): 93.89%
H^2 (unaccounted variability / sampling variability):   16.36
R^2 (amount of heterogeneity accounted for):            1.76%

Test for Residual Heterogeneity:
QE(df = 31) = 325.6335, p-val < .0001

Test of Moderators (coefficient(s) 2):
QM(df = 1) = 1.7815, p-val = 0.1820

Model Results:

         estimate      se     zval    pval    ci.lb   ci.ub
intrcpt    2.4016  0.2875   8.3541  <.0001   1.8382  2.9650  ***
task      -0.7720  0.5784  -1.3347  0.1820  -1.9056  0.3616

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3. A mixed-effects analysis with one binary moderator.

“split” function: “newd <- split(d, groupsize > -99)$‘TRUE’” splits off cases of *d* where the group size is greater than −99. We will also need to apply this to the variances and the group size variable itself: “newv <- split(v, groupsize > -99)$‘TRUE’;newgs <- split(groupsize, groupsize > -99)$‘TRUE’”. As is always the case, we should plot the relationship between continuous predictors before using continuous linear models. In *R*, this can be accomplished by entering “plot(newgs,newd,xlab=“Group Size”, ylab=“Effect Size”)”. See Figure 4 for the plot. There is one obvious large group, and most of the groups have *N* of three. Under such circumstances, the large observation is highly leveraged and could be unduly influential in the regression analysis; we might want to consider dichotomizing the variable (group size = 3 vs. group size > 3). However, the plot is not dramatically nonlinear, so we will begin with a continuous regression. We accomplish this by entering “rma(newd~newgs,newv)”; results appear in Figure 5.

The continuous predictor is vastly nonsignificant (z = 0.07, p = .95). The estimated variance component is now larger than it was in the simple random-effects model. This is counterintuitive, but not unheard of: the continuous predictor explains less heterogeneity than would be expected by random chance, so that the loss of degrees of freedom associated with estimating its slope has inflated the estimate of the variance component. (It is also not based on exactly the same data set, because of the listwise deletion of a case.) We may dichotomize the group size variable by creating a dummy indicator of size larger than 3: “sizedummy <- (newgs > 3)” creates a variable that has the value zero unless group size is larger than 3, where it has the value one. “rma(newd~sizedummy,newv)” also produces nonsignificant results (z = .74, p = .46).

The other continuous predictor we consider, total information load, also has missing values. We select cases of *d, v,* and *infoload* that have complete data, as in the previous analysis: “newd <- split(d,infoload > -99)$‘TRUE’; newv <- split(v,infoload > -99)$‘TRUE’; newinfoload <- split(infoload, infoload > -99)$‘TRUE’”. Once again, the scatterplot (“plot(newinfoload,newd,x lab=“Total Information Load”,ylab=“Effect Size”)”) shows an extreme value. We can conduct a random-effects meta-regression using “rma(newd~newinfoload,newv)”. By now, the

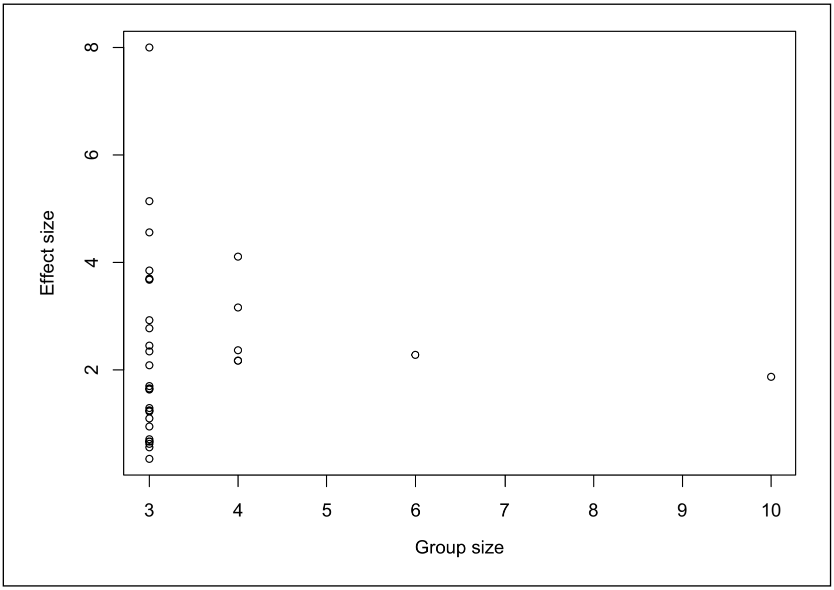
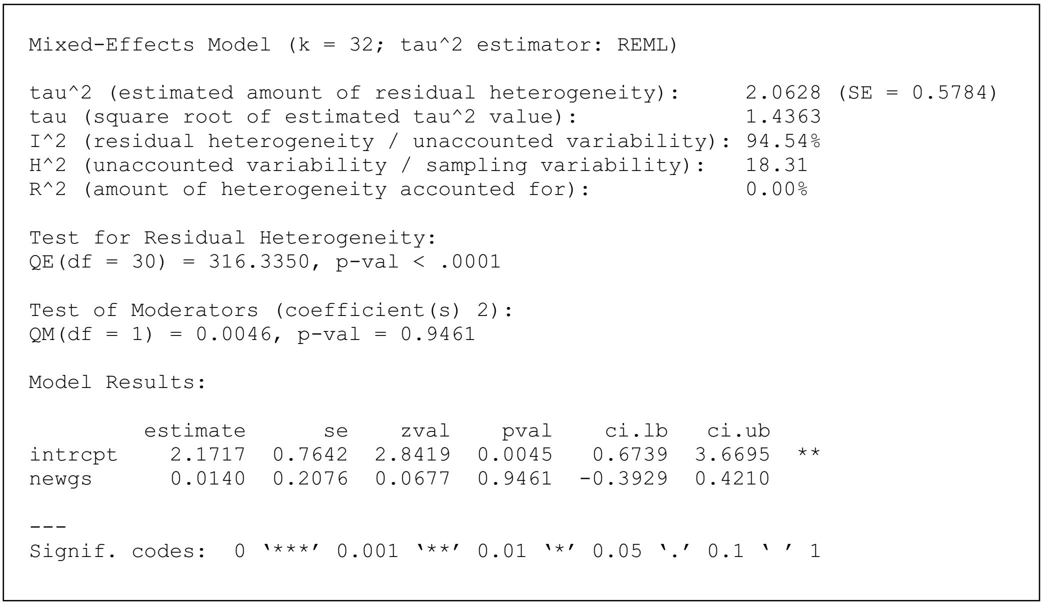


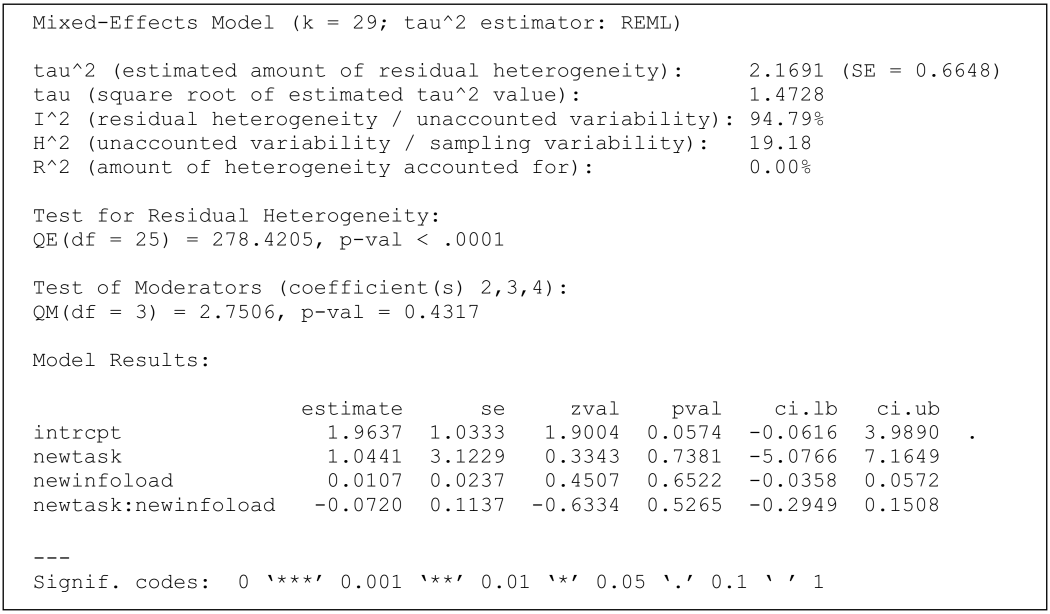
Figure 4. Plot of relation between effect size and group size.



```
Mixed-Effects Model (k = 32; tau^2 estimator: REML)

tau^2 (estimated amount of residual heterogeneity):     2.0628 (SE = 0.5784)
tau (square root of estimated tau^2 value):             1.4363
I^2 (residual heterogeneity / unaccounted variability): 94.54%
H^2 (unaccounted variability / sampling variability):   18.31
R^2 (amount of heterogeneity accounted for):            0.00%

Test for Residual Heterogeneity:
QE(df = 30) = 316.3350, p-val < .0001

Test of Moderators (coefficient(s) 2):
QM(df = 1) = 0.0046, p-val = 0.9461

Model Results:

         estimate      se    zval    pval    ci.lb   ci.ub
intrcpt    2.1717  0.7642  2.8419  0.0045   0.6739  3.6695  **
newgs      0.0140  0.2076  0.0677  0.9461  -0.3929  0.4210

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5. Analysis with a nonsignificant continuous predictor.

reader knows where to seek pertinent results in the *R* output, so we do not provide a figure here. The slope associated with information load is estimated to be .0018 and is nonsignificant (z = .77, p = .44). If we eliminate the outlying value where information load is 700 (" newd

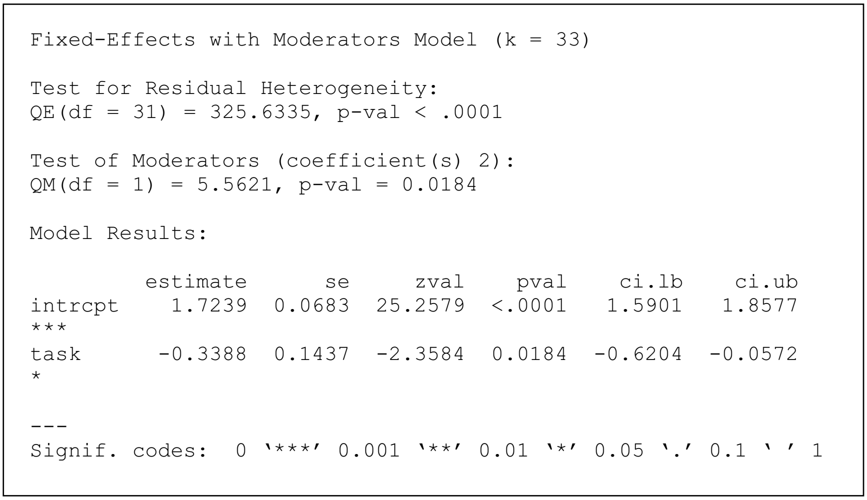


```
Mixed-Effects Model (k = 29; tau^2 estimator: REML)

tau^2 (estimated amount of residual heterogeneity):     2.1691 (SE = 0.6648)
tau (square root of estimated tau^2 value):             1.4728
I^2 (residual heterogeneity / unaccounted variability): 94.79%
H^2 (unaccounted variability / sampling variability):   19.18
R^2 (amount of heterogeneity accounted for):            0.00%

Test for Residual Heterogeneity:
QE(df = 25) = 278.4205, p-val < .0001

Test of Moderators (coefficient(s) 2,3,4):
QM(df = 3) = 2.7506, p-val = 0.4317

Model Results:

                     estimate      se     zval    pval    ci.lb   ci.ub
intrcpt                1.9637  1.0333   1.9004  0.0574  -0.0616  3.9890  .
newtask                1.0441  3.1229   0.3343  0.7381  -5.0766  7.1649
newinfoload            0.0107  0.0237   0.4507  0.6522  -0.0358  0.0572
newtask:newinfoload   -0.0720  0.1137  -0.6334  0.5265  -0.2949  0.1508

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6. Complex mixed-effects model with nonsignificant predictors.

<- split(newd,newinfoload < 700)$‘TRUE’; newv <- split(newv,newinfoload < 700)$‘TRUE’; newinfoload <- split(newinfoload,newinfoload < 700)$‘TRUE’”) and reestimate the model, the slope remains small (0.02) and nonsignificant (z = 1.00, p = .32).

Finally, we can consider models that incorporate multiple predictors simultaneously. One that makes particular sense here includes the categorical predictor *task type* and the continuous predictor *information load*, and allows for the possibility of an interaction between the two variables. The model effectively allows a separate meta-regression for each level of task type, with a common variance component. The formal model is given by $d = \beta_0 + \beta_1 I_{high} + \beta_3 Info + \beta_4 I_{high} \times Info$. Recalling that the I_{high} variable takes on only the values 0 and 1, this reduces to $d = \beta_0 + \beta_3 Info$ when the indicator is 0 and $d = (\beta_0 + \beta_1) + (\beta_3 + \beta_4) Info$ when the indicator is 1. We delete cases where the information load is missing or is an outlying value: “ newd <- split(d,(infoload > -99 & infoload < 700))$’TRUE’; newv <- split(v,(infoload > -99 & infoload < 700))$‘TRUE’; newinfoload <- split(infoload,(infoload > -99 & infoload < 700))$‘TRUE’; newtask <- split(task,(infoload > -99 & infoload < 700))$‘TRUE’”. Because we are estimating an interaction between a continuous and a dichotomous predictor, we do not center *information load*; centering in this case would not inflate standard errors or affect inference about the interaction, and it can actually reduce power for the dichotomous predictor. We estimate this more complex regression model: “rma(newd~newtask+newinfoload +newtask*newinfoload, newv)”. The results appear in Figure 6. Neither the predictors nor the interaction are individually significant. In addition, the output provides an overall test of significance for the entire model. This is denoted *QM*. Under the null hypothesis that the model as a whole does not explain variability in effect magnitude, it has a chi-square distribution with *df* equal to the number of predictors in the model. The overall model is nonsignificant (*QM* = 2.75, *df* = 3, p = .43).



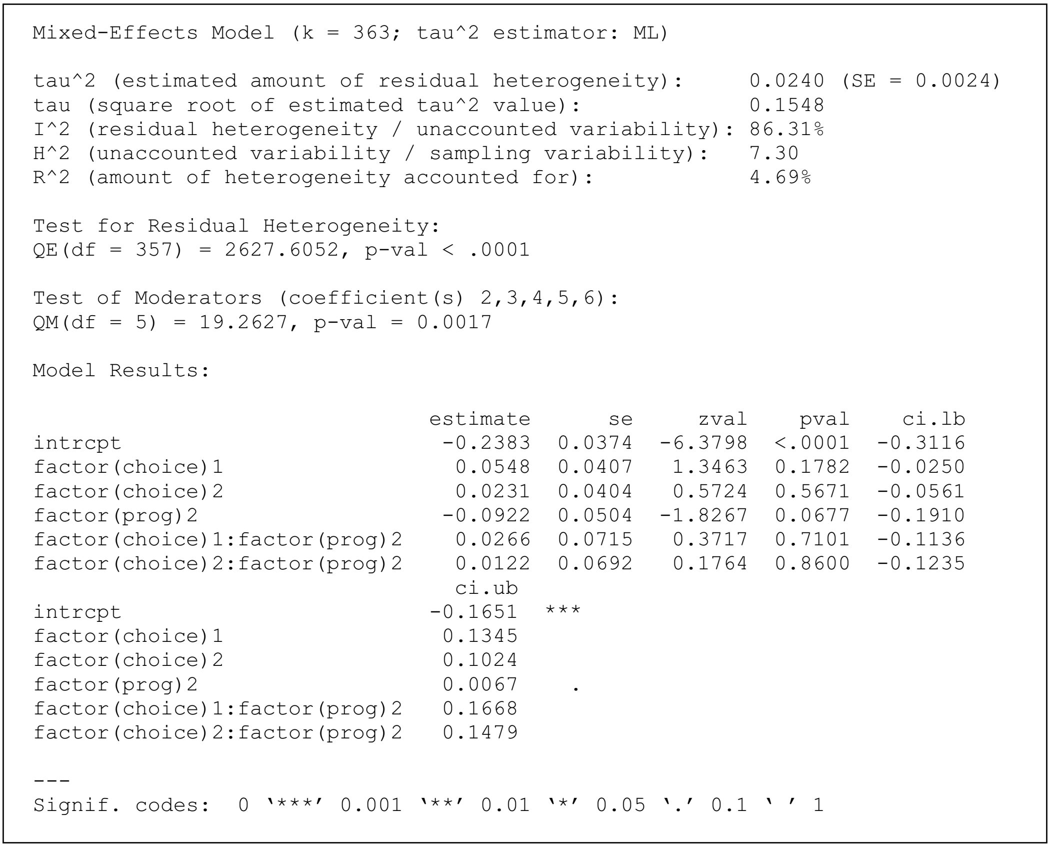
```
Fixed-Effects with Moderators Model (k = 33)

Test for Residual Heterogeneity:
QE(df = 31) = 325.6335, p-val < .0001

Test of Moderators (coefficient(s) 2):
QM(df = 1) = 5.5621, p-val = 0.0184

Model Results:

         estimate      se     zval    pval    ci.lb    ci.ub
intrcpt    1.7239  0.0683  25.2579  <.0001   1.5901   1.8577
***
task      -0.3388  0.1437  -2.3584  0.0184  -0.6204  -0.0572
*

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7. A fixed-effects analysis with one binary moderator.

Conditionally random inference. The analyst who decides to make the choice of fixed- or random-effects models conditional on heterogeneity will begin the moderator analyses using fixed-effects inference. A fixed-effects analysis of the first model, which considers task type as a moderator of effect size, is given by “rma(d~task,v, method=‘FE’)”; results appear in Figure 7. Now, the distinction between high and low demonstrability tasks appears significant ($z = -2.3584$, $p = .0184$). However, the Q test for residual heterogeneity ($QE = 325.63$, $df = 31$, $p < .0001$), is highly significant. Hence, the analyst approaching model selection conditional on the heterogeneity test would conclude that the fixed-effects model does not fit, and would move to the mixed-effects model that we saw in the previous section (in which task type was nonsignificant). Similar findings hold for the other moderator analyses we considered before.

The fixed-effects test for group size (whether continuous or dichotomized) is highly significant. For example, the dichotomized group size analysis results in an increment of 1.07 to the effect size for group sizes greater than 3 ($z = 6.40$, $p < .0001$). Once again, though, heterogeneity is highly significant ($QE = 288.18$, $df = 30$, $p < .0001$); so again, the responsible analyst would conclude that the fixed-effects model does not fit, and would estimate the mixed-effects model, where the effect of group size was nonsignificant. Similar findings occur with the other moderators: all are significant in the fixed-effects analysis, but there is always significant residual heterogeneity, which demands return to the mixed-effects model where the moderators are nonsignificant. Hence, conditionally random inference for this data set ends up being identical to mixed-effects inference.

The reader may well wonder why this is happening. The fact is, in highly heterogeneous data sets such as this example, almost any potential moderator can be found significant in a fixed-effects analysis, regardless of whether it is systematically associated with effect size. To illustrate this point, consider a randomly generated sequence of zeros and ones, representing a potential moderator that is clearly not systematically related to effect size. In *R*, we can generate such a set of values by entering “x <- rbinom(33, 1, .5)”. This simulates a set of 33 flips of a coin, with “heads” recorded as 1 and “tails” as 0. A fixed-effects meta-analysis using this randomly generated dummy variable may be conducted using “rma(d~x,v)”. The reader who endeavors to try this repeatedly will find that for this data set, the



```
Mixed-Effects Model (k = 363; tau^2 estimator: ML)

tau^2 (estimated amount of residual heterogeneity):     0.0240 (SE = 0.0024)
tau (square root of estimated tau^2 value):             0.1548
I^2 (residual heterogeneity / unaccounted variability): 86.31%
H^2 (unaccounted variability / sampling variability):   7.30
R^2 (amount of heterogeneity accounted for):            4.69%

Test for Residual Heterogeneity:
QE(df = 357) = 2627.6052, p-val < .0001

Test of Moderators (coefficient(s) 2,3,4,5,6):
QM(df = 5) = 19.2627, p-val = 0.0017

Model Results:

                               estimate      se     zval    pval    ci.lb
intrcpt                         -0.2383  0.0374  -6.3798  <.0001  -0.3116
factor(choice)1                  0.0548  0.0407   1.3463  0.1782  -0.0250
factor(choice)2                  0.0231  0.0404   0.5724  0.5671  -0.0561
factor(prog)2                   -0.0922  0.0504  -1.8267  0.0677  -0.1910
factor(choice)1:factor(prog)2    0.0266  0.0715   0.3717  0.7101  -0.1136
factor(choice)2:factor(prog)2    0.0122  0.0692   0.1764  0.8600  -0.1235
                                  ci.ub
intrcpt                         -0.1651  ***
factor(choice)1                  0.1345
factor(choice)2                  0.1024
factor(prog)2                    0.0067    .
factor(choice)1:factor(prog)2    0.1668
factor(choice)2:factor(prog)2    0.1479

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 8. A mixed-effects analysis with an interaction.

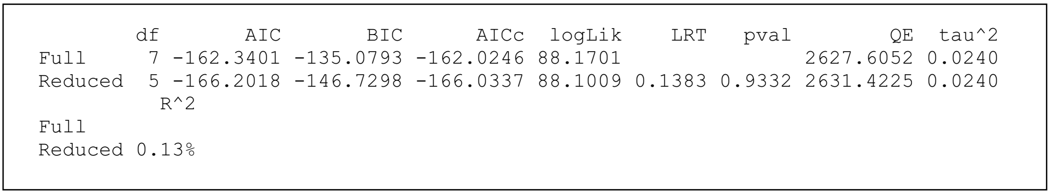
moderator will be found significant at the .05 level or better more than half of the time. However, 100% of these analyses will be found to be heterogeneous at the .05 level, which will again require a return to the mixed-effects model, where the predictor will be nonsignificant. (These assertions are based on a simulation with 10,000 replications of a randomly generated moderator using the Li Lu et al. (2012) data. Recall that our analyses are not direct replications of those presented in the Li Lu et al. paper, and are not intended to be interpreted as such.)

The question is, how willing will we be to give up an effect that makes theoretical sense and is significant in a fixed-effects model but shows significant heterogeneity? The psychology literature has many examples of meta-analyses that report significant fixed-effects moderators, but fail to report on residual heterogeneity. We wonder how many of these findings may represent spurious results.

Pettigrew and Tropp (2006)

Next we demonstrate some analyses using a much larger dataset from a meta-analysis conducted by Pettigrew and Tropp (2006). We will use these data to explore likelihood ratio testing in the context of mixed-effects meta-analysis. We will also discuss how to create a funnel plot and conduct some sensitivity analyses to assess the presence of publication bias.

This meta-analysis examined the relationship between intergroup contact and intergroup prejudice. Studies have concluded that contact between groups reduces prejudice; however,



```
        df       AIC       BIC      AICc  logLik    LRT   pval        QE  tau^2
Full     7 -162.3401 -135.0793 -162.0246 88.1701               2627.6052 0.0240
Reduced  5 -166.2018 -146.7298 -166.0337 88.1009 0.1383 0.9332 2631.4225 0.0240
          R^2
Full
Reduced 0.13%
```

Figure 9. A likelihood-ratio test of the interaction.

many aspects of this relationship are unclear. Some find that contact has to occur under optimal conditions, and unfavorable conditions can increase prejudice (Pettigrew & Tropp, 2006). Other research indicates that characteristics of the setting and types of groups can alter the effects of contact. Pettigrew and Tropp (2006) seek to clarify those results. The original dataset consists of 713 correlation coefficients (*r*) from independent samples in 515 individual studies. Each effect size represents a correlation between intergroup contact and prejudice. The correlations were transformed by Fisher's *Z* ($Z = 1/2 \log((1+r)/(1-r))$), a common transformation that results in a more normal sampling distribution with sampling variance equal to $1/(N - 3)$. To make the dataset more manageable, we arbitrarily truncate the original sample to 363 effects.[4]

The authors code 14 moderators. Some assess research quality, including source of publication (published or unpublished), type of design (within- or between-subjects), type of study (survey, quasi-experiment, or experiment), type of control group (whether they have no, some, or considerable earlier contact with the outgroup), type of contact measure (whether researchers assume contact had occurred, observe it, or report it), and quality of measures for contact and prejudice. One study characteristic that they consider of particular interest is the question of whether there was a program in the study that was structured according to Allport's (1954) theory. They also code participant characteristics, including age, sex, geographical area of the study, and kind of target group (e.g., sexual orientation, physically disabled, mentally ill, etc.). They also code moderators about the contact, including its setting (e.g., laboratory, recreational, etc.), and whether participants have no choice, some choice, or full choice to participate in the contact. All of these potential moderators are categorical variables.

Mixed-effects inference. We use a small subset of those variables: choice, publication status, and program structure (no program, or a structured program). In our reduced data set, there are no missing data. We read the data into *R* and attach the dataset: “Pettigrew <- read.table(“path/to/file/pettigrew.txt”, col.names=c(“z”,“N”,“choice ”,“pub”,”prog“)); attach(Pettigrew)”. This dataset has *N* rather than sampling variance, so we need to calculate the variances: “v <- 1/(N-3)”. It is useful to have a sense of the frequency of the various levels of the moderators. We can accomplish this by creating tables: “table(choice); table(pub); table(prog)”. The command reveals that there are 62 effects from studies with no choice (choice = 0), 139 with some choice (choice = 1), and 162 with full choice (choice = 2); there are 296 published effects and 67 unpublished; and there are 282 effects with no program and 81 from studies with a structured program. Note that the first of these variables is a factor with three levels. The other two have only two levels (like the binary predictor in the Li Lu et al. (2012) example), but the levels are coded “1” and “2” rather than “0” and “1,” so we will not be able to include them as predictors as if they were dummy variables.

We focus here on choice and structure; we will consider publication status later in a sensitivity analysis. As in the other example, we begin with random-effects inference. This time, however, we will take a reductionist approach, starting with the most complex model, and assess whether we can reduce the model. The *metafor* package has the ability to develop coding for ANOVA-like factors; all we need to do is specify that a variable is a factor, using “factor(VarName)” (where “VarName” is the name of the variable). The most complex model here would be an ANOVA-like analysis that includes main effects for choice and program structure, along with an interaction between those factors. This can be accomplished with the following command: “rma(z~factor(choice) + factor(prog) + factor(choice)*factor(prog), v, method=‘ML’)”. The results appear in Figure 8.

We notice first that even with these moderators, there is residual heterogeneity. The variance component is estimated to be .024, and is significantly different from zero (QE = 2627.61, df = 357, p < .0001); the corresponding percentage of total variation that is between-effects (I^2) is 86. We also see that the model as a whole is significant (QM = 19.26, df = 5, p = .0017), indicating that some combination of choice, program type, and the interaction is associated with variation in effect magnitude. However, without understanding how *metafor* parameterizes the problem, it is difficult to make sense of the detailed output. The situation is analogous to a manually coded ANOVA: individual coefficients collectively make up effects, but a collective test is needed to test concepts like interaction. We can see that the last two coefficients make up the interaction and are individually nonsignificant; but are the two coefficients simultaneously significant? A 2 *df* test for the interaction is needed.

The astute reader will have noticed a difference in our use of *rma* this time: we specified “method=‘ML’”. When no method is specified, *rma* estimates the variance component using a method known as “restricted maximum likelihood.” This approach corrects for bias in the unrestricted “method=‘ML’” estimate. Ordinarily, that is a good thing. Here, however, we will use an approach known as *likelihood-ratio testing* to perform hypothesis tests about groups of parameters that make up effects like the interaction. In order to meet the assumptions of likelihood ratio testing, we need to use the true maximum likelihood estimate of the variance component.

Likelihood-ratio testing is a general method for comparing nested models. One model is said to be nested within another if one can change the more complex model into the simpler model by fixing or constraining parameters. Twice the difference in the log-likelihoods of the two models has a chi-square distribution with *df* equal to the number of fixed or constrained parameters if the simpler model is as good as the complex model. Here, for example, we can compare the model with the interaction to a model with only main effects by fixing the two parameters that make up the interaction to be zero; that is exactly equivalent to estimating a main-effects-only model.

The *metafor* package provides a simple way to do this test. If we save the results of a call to *rma* as a complex variable, and then fit and save a nested model, an *anova* function will compare the two models and calculate the likelihood ratio test. To test the interaction, we estimate and save the model with interaction, arbitrarily choosing the name “model1” for those results: “ model1 <- rma(z~factor(choice) + factor(prog) + factor(choice)*factor(prog), v, method=‘ML’)” . Then we estimate and save a simpler model that does not include the interaction: “ model2 <- rma(z ~ factor(choice) + factor(prog), v, method=‘ML’)”. Finally, we compare the two models: “anova(model1, model2)”. The results of this comparison appear in Figure 9. The likelihood-ratio chi-square has the value 0.1383. Its *df* is equal to the difference in degrees of freedom for the two models: 7 − 5 = 2. It is vastly nonsignificant (p = .9332).

If we look at the results of the main-effects model (without interaction) by typing “model2”, we see that the *program* factor (which has only two levels and hence one variable in the model) is significant: z = −2.78, p = .0055. However, the information about the three-level *choice* factor is ambiguous: one of the two variables that make up this factor is significant (z = 1.97, p = .0487), but the other is not (z = 0.95, p = .3416). So is the *choice* factor significant? We can test this using another likelihood-ratio test: compare the main-effects model (model2) with a model that includes only the *program* factor: "model3 <- rma(z ~ factor(prog), v, method='ML'); anova(model2, model3)". The resulting likelihood-ratio chi-square statistic is marginally significant ($\chi^2 = 4.72$, $df = 2$, $p = .0945$). So the only significant predictor is the *program* variable.

Conditionally random inference. We can conduct the likelihood-ratio test of the interaction using fixed-effects models by repeating the estimation of "model1" and "model2," this time specifying "method='FE'" instead of "method='ML'", and comparing the results using "anova(model1, model2)". The resulting likelihood-ratio test is nonsignificant ($\chi^2 = 3.82$, $df = 2$, $p = .1483$). If we go on to compare the main-effects model with a model that includes only the *program* factor using fixed-effects inference ("model3 <- rma(z ~ factor(prog), v, method='FE'); anova(model2, model3)"), the likelihood-ratio test for the *choice* effect is highly significant ($\chi^2 = 152.11$, $df = 2$, $p < .0001$). However, there is significant residual heterogeneity in the model ($QE = 2631.42$, $df = 359$, $p < .0001$), so the analyst using conditionally random inference would reject that model and move to mixed-effects inference, reaching the same conclusion as in the a priori mixed-effects analysis. Once again, then, conditionally random inference leads to the same findings, but requires that the analyst be willing to reverse a decision about a significant finding because of the presence of heterogeneity.

Sensitivity analysis for publication bias. We discuss sensitivity analyses and various approaches to assessing publication bias in the next section. (Publication bias is the inflation of effect size estimates that occurs because larger or more statistically significant effects are more likely to be published.) We illustrate one frequently used approach here: the comparison of effects from published and unpublished studies. The approach is not without shortcomings; in particular, published and unpublished studies may differ on other criteria such as methodological quality. Nevertheless, the comparison can be informative. A mixed-effects model comparing published and unpublished studies is given by "rma(z~pub,v)". The results show that effect sizes in published studies are actually about 0.05 *smaller* than effects from unpublished studies (although the finding is not significant: $z = -1.76$, $p = .0782$). Hence, the analysis does not suggest that publication bias is a likely problem here.

Bells, Whistles, and Sensitivity Analyses

The *metafor* package has many other features that limited space prevents our covering here. We mention a few briefly.

Knapp and Hartung (2003) describe a method for addressing uncertainty associated with the estimation of the variance component. Ordinarily, inference about single coefficients in a meta-analytic model relies on large sample theory z tests. These tests do not fully recognize sampling uncertainty of the variance component. Knapp and Hartung's (2003) t approximation compensates for that uncertainty. This method may be accessed in any random or mixed-effects analysis by specifying "knha=TRUE" when using *rma.*

Many practitioners find forest plots useful for smaller meta-analyses. The forest plot graphically presents a confidence interval for each individual effect estimate. These may be performed in *metafor* by specifying "forest(effect, variance)" (where "effect" and "variance" are the names of the effect size variable and the sampling variance variable).

Often, researchers will want to have model-predicted conditional mean effects, with standard errors (and perhaps confidence intervals). The *metafor* package provides a means of accessing these. Given saved output from an *rma* analysis called "mymodel," "predict(mymodel)" will produce conditional mean effect magnitudes with standard errors and confidence intervals for all cases in the dataset. If continuous predictors are involved in the model, these conditional means will potentially be different for every case. However, when all of the predictors in the model are categorical, any case having a particular combination of moderator values will have the same predicted mean, and the output can be taken as the estimated effect for studies with that combination of characteristics.

No meta-analysis would be complete without careful sensitivity analyses. We have already seen an example when we reestimated the *group size* effect in the Li Lu et al. (2012) example, categorizing the size to reduce the influence of the extremely large group. We also assessed the effect of *information load* without the extreme value. Other sensitivity analyses that are frequently performed involve putting a floor on the sampling variance of an individual effect to reduce the influence of very large studies in a meta-analysis. This can easily be accomplished by setting the variances equal to the maximum of the actual sampling variances and the floor. For example, if we wanted to prohibit sampling variances smaller than .001 (which is roughly the variance of a Fisher *z* value with a sample size of 1,000), we could specify "newv <- max(.001, v)" and use *newv* in place of *v* when we use *rma*.

A particularly important subject for sensitivity analysis is the possibility of publication bias; no meta-analysis may be considered complete without some consideration of this issue. One method for addressing this is the *funnel plot* (Light & Pillemer, 1984; Sterne & Egger, 2001). Funnel plots are available in *metafor* using the *funnel* function: "funnel(mymodel)". Another approach to assessing publication bias involves regressions of effect size on sampling uncertainty. This approach, often referred to as "Egger regression" (Egger, Smith, Schneider, & Minder, 1997) is available in *metafor* through the *regtest* function. Yet another method, *trim-and-fill* (Duvall & Tweedie, 2000a, 2000b) is available through the *trimfill* function.

Concluding Remarks

We have seen that the method of maximum likelihood provides a versatile approach to estimation and inference for complex meta-analytic models. The *R* package *metafor* makes this method easily and freely available to statistical consumers who have reasonable levels of sophistication; with a little experience, estimating these models is not more difficult than conducting ANOVA and regression analyses for primary data.

We have advocated an a priori choice of random- and mixed-effects models for assessing moderators of effect size, based on an argument that the goals of meta-analytic inference demand that choice. However, we acknowledge that a conditionally random approach in which the choice of fixed- or random-effects models is decided by a test of residual heterogeneity is ubiquitous in practice, in part because most manuals for the conduct of meta-analysis present that approach. We have seen that under representative circumstances the a priori random and conditionally random approaches often result in the same final conclusions about the significance of moderators. However, it is important to remember that the conditionally random approach often requires that the analyst have the self-discipline to give up on an apparently significant moderator because its significance does not persist in the context of a mixed-effects model. We believe (but cannot prove) that the meta-analytic literature in psychology contains many cases in which that self-discipline was lacking.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

[1] We follow the convention of using Courier font whenever text represents a command typed in *R*.

[2] Note that different effect types (e.g., mean differences and correlations) should not be combined in a single meta-analysis without an equivalence transformation. Formulas exist for conversion between effect-size metrics; see, for example, Borenstein (2009).

[3] Please refer the Online Supplementary Material for the dataset.

[4] Please refer the Online Supplementary Material for the dataset.

References

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison Wesley.

Anderson, C., Shibuya, A., Ihori, N., Swing, E., Bushman, B., Sakamota, A., … Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: A meta-analytic review. *Psychological Bulletin*, *136*, 151–173. doi:10.1037/a0018251

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 357–381). New York, NY: Russell Sage Foundation.

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods*, *14*, 225–238. doi:10.1037/a0016619

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 279–293). New York, NY: Russell Sage Foundation.

Burnstein, E., & Vinokur, A. (1977). Persuasive argumentation and social comparison as determinants of attitude polarization. *Journal of Experimental Social Psychology*, *13*, 315–332. doi:10.1016/0022-1031(77)90002-6

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

Cooper, H., Hedges, L.V., & Valentine, J. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russel Sage Foundation.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*, 177–188. doi:10.1016/0197-2456(86)90046-2

Duvall, S. J., & Tweedie, R. L. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89–98.

Duvall, S. J., & Tweedie, R. L. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463. Retrieved from http://www.jstor.org/stable/2676988

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple graphical test. *British Medical Journal*, *315*, 629–634. doi:10.1136/bmj.315.7109.629

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *10*, 3–8. Retrieved from http://www.jstor.org/stable/1174772

Hartung, J., Knapp, G., & Sinha, B. K. (2008). *Statistical meta-analysis with applications*. Hoboken, NJ: John Wiley & Sons.

Hattie, J., Rogers, H. J., & Swaminathan, H. (2014). The role of meta-analysis in educational research. In A. Reid, E. P. Hart, & M. Peters (Eds.), *A companion to research in education* (pp. 197–207). Dordrecht, Netherlands: Springer Science and Businesses Media.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–299). New York, NY: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504. doi:10.1037/1082-989X.3.4.486

Knapp, G., & Hartung, J. (2003). Improved tests for a random-effects meta-regression with a single covariate. *Statistics in Medicine*, *22*, 2693–2710. doi:10.1002/sim.1482

Konstantopoulos, S., & Hedges, L. V. (2009). Analyzing effect sizes: Fixed-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 279–293). New York, NY: Russell Sage Foundation.

Koricheva, J., Gurevitch, J., & Mengersen, K. (Eds.). (2013). *Handbook of meta-analysis in ecology and evolution*. Princeton, NJ: Princeton University Press.

Leandro, G. (2007). *Meta-analysis in medical research: The handbook for the understanding and practice of meta-analysis*. Oxford, UK: Blackwell.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Li Lu, J., Yuan, Y.C., & McLeod, P. L. (2012). Twenty-five years of hidden profiles in group decision making: A meta-analysis. *Personality and Social Psychology Review*, *16*, 54–75. doi:10.1177/1088868311417243

Pettigrew, T., & Tropp, L. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*, 751–783. doi:10.1037/0022–3514.90.5.751

Pigott, T. (2009). Handling missing data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 399–416). New York, NY: Russell Sage Foundation.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*, 97–129. doi:10.1348/000711007X255327

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York, NY: Russell Sage Foundation.

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 257–277). New York, NY: Russell Sage Foundation.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, *54*, 1046–1055. doi:10.1016/S0895-4356(01)00377-8

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research*. Chichester, UK: John Wiley & Sons.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/

Wang, M. C., & Bushman, B. J. (1999). *Integrating results though meta-analytic review using SAS software*. Cary, NC: SAS Institute, Inc.

Wildschut, T., Pinter, B., Vevea, J. L., Insko, C. A., & Schopler, J. (2003). Beyond the group mind: A quantitative review of the interindividual-intergroup discontinuity effect. *Psychological Bulletin*, *129*, 698–722. doi:10.1037/0033-2909.129.5.698

Wilson, D. (2001). Meta-analytic methods for criminology. *Annals of the American Academy of Political and Social Science*, *578*, 71–89. doi:10.1177/000271620157800105