Received: 20 November 2020 | Revised: 18 March 2021 | Accepted: 22 April 2021

DOI: 10.1002/jrsm.1489

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Using information-theoretic approaches for model selection in meta-analysis

Ozan Cinar[1] | James Umbanhowar[2] | Jason D. Hoeksema[3] | Wolfgang Viechtbauer[1]

[1]Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, The Netherlands

[2]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[3]Department of Biology, University of Mississippi, University, Mississippi, USA

Correspondence
Ozan Cinar, Department of Psychiatry and Neuropsychology, Maastricht University, Vijverdalseweg 1, Maastricht 6226 NB, The Netherlands.
Email: ozan.cinar@maastrichtuniversity.nl

Meta-regression can be used to examine the association between effect size estimates and the characteristics of the studies included in a meta-analysis using regression-type methods. By searching for those characteristics (i.e., moderators) that are related to the effect sizes, we seek to identify a model that represents the best approximation to the underlying data generating mechanism. Model selection via testing, either through a series of univariate models or a model including all moderators, is the most commonly used approach for this purpose. Here, we describe alternative model selection methods based on information criteria, multimodel inference, and relative variable importance. We demonstrate their application using an illustrative example and present results from a simulation study to compare the performance of the various model selection methods for identifying the true model across a wide variety of conditions. Whether information-theoretic approaches can also be used not only in combination with maximum likelihood (ML) but also restricted maximum likelihood (REML) estimation was also examined. The results indicate that the conventional methods for model selection may be outperformed by information-theoretic approaches. The latter are more often among the set of best methods across all of the conditions simulated and can have higher probabilities for identifying the true model under particular scenarios. Moreover, their performance based on REML estimation was either very similar to that from ML estimation or at times even better depending on how exactly the REML likelihood was computed. These results suggest that alternative model selection methods should be more widely applied in meta-regression.

KEYWORDS

information criteria, meta-analysis, meta-regression, model selection, multimodel inference

1 | INTRODUCTION

Empirical research often involves taking measurements to assess the magnitude of a treatment effect, the size of a group difference, or the direction and strength of the relationship between two variables (i.e., an "effect size"). When multiple estimates of a phenomenon of interest are available from a collection of studies, meta-analytic methods can be used to synthesize the estimates into an overall value that typically will be more precise than the individual estimates.[1] In addition, since the aggregated value is usually a reflection of estimates obtained under varying conditions and circumstances, it carries with it a sense of greater generalizability.

At the same time, those varying conditions and circumstances can also lead to variability in the underlying true effects, a phenomenon typically referred to as "heterogeneity."[2] In this case, the observed estimates will tend to be more variable than would be expected if the true effects were identical across studies (i.e., under homogeneity). When such heterogeneity is detected, attempts are often made to find its potential sources. For this, one can examine whether the effect size estimates are systematically related to the conditions and circumstances under which they were obtained. Such moderator analyses are commonly conducted by means of an approach called meta-regression.[3-7] Here, one first codes the various (study-level) characteristics of interest into a set of predictor variables or "moderators." Next, regression-type models are used to examine the relationship between moderators and the estimates.

In practice, meta-analytic datasets tend to be highly multi-factorial in nature, containing a large number of study characteristics that are potentially relevant and plausible predictors of the observed effects.[7] Moreover, such analyses are inherently observational in nature because the values of the explanatory variables have not been independently/systematically manipulated.[7] As a result, the various study characteristics will often be correlated with each other, complicating the determination of the unique contribution of individual moderators to the heterogeneity.

Nevertheless, it is still common practice to examine one moderator variable at a time by means of a series of univariate or "single-factor" meta-regression models.[8] However, unless steps are taken to control the familywise Type I error rate (which is not common practice in this context), doing so leads to a high chance of false positive findings. Moreover, due to their correlation, moderator variables found to be relevant predictors are likely to account for shared variability in the effect sizes. Fitting meta-regression models containing multiple predictors of interest may circumvent the latter issue, but it is unclear how moderators should be selected for inclusion in such multi-factor models. For example, stepwise procedures have repeatedly been argued to be of limited value.[9,10] Also, parameter estimates and the statistical significance of particular moderators may vary substantially among models, depending on which other moderators are included.

Although typically not framed in this manner, fitting meta-regression models is in essence a form of model selection. Accordingly, one could also consider to make use of methods for model selection and multimodel inference based on information-theoretic approaches.[9,11,12] However, it is far from straightforward to apply information criteria in model selection in this context. To start, it is not clear which estimation procedure to use for model fitting (e.g., ML, REML), nor which information criterion to choose (e.g., AIC, BIC, AICc). On the one hand, it has been argued that information criteria under REML (restricted maximum likelihood) estimation should not be used to compare models differing in their fixed effects[13,14] because the contrasts used to derive the restricted likelihood (which is then used in the calculation of the information criteria) depend on the model matrix of the fixed effects. However, simulation results from Gurka[15] suggest that this problem may not be substantial for some types of models and datasets. Moreover, we might prefer to use REML over ML (maximum likelihood) estimation, since the latter may produce biased parameter estimates (especially for variance components in the model), which in turn affect the calculation of the information criteria.[16] We are aware of two simulation studies exploring these issues,[15,17] but results from those studies are inconclusive and only narrowly related to the models and data structures used in meta-analyses.

The purpose of this paper is therefore to examine methods for model selection when faced with multifactorial data in the meta-analytic context. We are specifically interested in how well methods based on information criteria and multimodel inference compare against simpler methods such as univariate testing of individual moderator variables or fitting a single model including all potential moderator variables of interest. To examine these issues, we conducted a simulation study comparing the various approaches using data reflective of a wide variety of circumstances one may encounter in practice.

The outline of the paper is as follows. In Section 2, we describe the meta-analytic random- and mixed-effects meta-regression models, ML and REML estimation thereof, and standard methods for testing coefficients in the context of these models. In Section 3, model selection via testing, information criteria, multimodel inference, and relative variable importance are described. In Section 4, we then provide an example illustrating the various approaches based on a dataset containing the results from studies examining the effects of inoculation with root-symbiotic mycorrhizal fungi on plant biomass. The methods and results for the simulation study comparing the various approaches are described in Section 5. We then conclude the paper with a discussion of the findings in Section 6, where we address some additional issues and concerns.

2 | META-ANALYSIS MODELS

We assume that k independent studies have been selected for inclusion in a meta-analysis and that each study provides a single effect size estimate or observed outcome. For example, for a set of studies examining the effectiveness of a particular experimental treatment, the outcome measure may be the raw or standardized mean difference or log response ratio.[18] When examining the relationship between two variables, the outcome measure may be the raw or Fisher's r-to-z transformed correlation coefficient.[18] In the health/medical sciences, dependent variables are often measured dichotomously, leading to (log-transformed) odds/risk ratios and risk differences as effect size measures of choice.[19]

Regardless of the outcome measure used, let y_i denote the observed value in the ith study and θ_i the corresponding true parameter. For brevity, we will refer to y_i as the effect size estimate and θ_i as the true effect size. We assume that

$$y_i = \theta_i + \varepsilon_i, \tag{1}$$

where $\varepsilon_i \sim N(0, v_i)$ denotes the sampling error and v_i the sampling variance of the ith estimate. For all of the effect size measures commonly used in meta-analysis (and all of the measures noted above), we can derive an estimate of v_i for each study.[18,19] The sampling variances are typically treated as known constants in the analyses (although technically some uncertainty is still attached to them, especially in small samples).

2.1 | Random- and mixed-effects models

According to the random-effects model,[5] the true effect sizes are heterogeneous and are given by

$$\theta_i = \mu + u_i, \tag{2}$$

where $u_i \sim N(0, \tau^2)$. Therefore, τ^2 denotes the amount of heterogeneity in the true effects and $E[\theta_i] = \mu$ the average true effect. A special case of the random-effects model arises when $\tau^2 = 0$, in which case the true effects are homogeneous.

Heterogeneity in the true effects is often not purely random, as assumed by the random-effects model, but a result of systematic differences between the studies (e.g., in terms of how an experimental treatment was implemented). Assume that information about p potential moderator variables has been extracted from the studies along with the effect size estimates. We can then set up a mixed-effects meta-regression model of the form

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + u_i, \tag{3}$$

where x_{ij} denotes the observed value of the jth moderator variable in the ith study, β_j ($j = 1, \ldots, p$) denotes how $E[\theta_i]$ changes for a one-unit increase in x_{ij}, and $u_i \sim N(0, \tau^2)$ as before, but τ^2 now denotes residual heterogeneity, that is, variability in the true effects not accounted for by the moderators included in the model.[3-7]

The meta-regression model above can accommodate (a mixture of) quantitative and qualitative moderator variables (the latter through appropriate dummy coding of the various levels of the factor) and x_{ij} may also reflect an interaction term between two or more moderator variables or polynomial/spline functions of individual moderators (to model the non-linear influence of a quantitative moderator variable on the effect sizes). However, we will only consider main effects throughout this paper, as models involving higher-order terms are not frequently used in practice.

2.2 | Model fitting and inference

Let $\boldsymbol{X}$ denote the $(k \times (p+1))$ model matrix containing the values of the p moderator variables with a vector of ones in the first column, corresponding to the model intercept. In fact, the random-effects model is just a special case of the mixed-effects model, where $\boldsymbol{X}$ simply consists of a column of ones. Next, let $\boldsymbol{y}$ denote the $(k \times 1)$ vector with the observed effect size estimates and $\boldsymbol{V}$ a $(k \times k)$ diagonal matrix with the sampling variances (i.e., the v_i values) along the diagonal. The random/mixed-effects model can then be written as

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{M}), \tag{4}$$

where $\boldsymbol{M} = \boldsymbol{V} + \tau^2 \boldsymbol{I}$ and $\boldsymbol{I}$ denotes a $(k \times k)$ identity matrix. Letting $\boldsymbol{W} = \boldsymbol{M}^{-1}$, the log likelihood function is therefore given by

$$\begin{aligned} ll_{\mathrm{ML}}(\boldsymbol{\beta}, \tau^2) = & -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{M}| \\ & -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X\beta}), \end{aligned} \tag{5}$$

which depends on $\boldsymbol{\beta}$ and τ^2. For a given value of τ^2, the maximum likelihood estimate of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}. \tag{6}$$

Hence, finding the maximum likelihood estimates of $\boldsymbol{\beta}$ and τ^2 is considerably simplified by maximizing the profiled log likelihood

$$ll_{\mathrm{ML}}(\tau^2) = -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{M}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})'\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}) \tag{7}$$

over τ^2 and then obtaining the maximum likelihood estimates of the elements in $\boldsymbol{\beta}$ with (6).

Maximum likelihood estimates of τ^2 are known to be negatively biased in small samples (i.e., when k is small).[20] On the other hand, restricted maximum likelihood (REML) estimation yields (approximately) unbiased estimates and is therefore to be preferred when unbiasedness is deemed important. The restricted log likelihood function is given by

$$\begin{aligned} ll_{\mathrm{REML}}(\tau^2) &= -\frac{k-p-1}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{X}'\boldsymbol{X}| - \frac{1}{2}\ln|\boldsymbol{M}| \\ &\quad -\frac{1}{2}\ln|\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})'\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}). \end{aligned} \tag{8}$$

Since the REML likelihood only depends on τ^2, maximization of $ll_{\mathrm{REML}}(\tau^2)$ is again a one-dimensional optimization problem. Once the REML estimate of τ^2 has been obtained, we can again estimate the elements in $\boldsymbol{\beta}$ with (6).

Maximization of (7) and (8) can be easily accomplished either by an exhaustive search or by means of an optimization algorithm, such as gradient ascent or some Newton-type algorithm.[21] Below, we will denote the maximized values of (7) and (8) as ll_{ML} and ll_{REML}, respectively, and the estimate of τ^2 that is obtained with either method by $\hat{\tau}^2$.

Once τ^2 has been estimated with either ML or REML estimation, the estimated model coefficients (i.e., b_0, b_1, ..., b_p) are then given by (6), with $\hat{\tau}^2$ substituted for τ^2 in $\boldsymbol{M}$ and hence $\boldsymbol{W}$. The variance–covariance matrix of the model coefficients can then be estimated with

$$\mathrm{Var}[\boldsymbol{b}] = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}. \tag{9}$$

The diagonal elements of $\mathrm{Var}[\boldsymbol{b}]$ are the estimated sampling variances of the model coefficients (i.e., $\mathrm{Var}[b_0]$, $\mathrm{Var}[b_1]$, ..., $\mathrm{Var}[b_p]$). Taking the square-root thereof provides the estimated standard errors (i.e., $\mathrm{SE}[b_0]$, $\mathrm{SE}[b_1]$, ..., $\mathrm{SE}[b_p]$). A Wald-type test of an individual moderator variable can then be conducted by comparing

$$z_j = \frac{b_j}{\mathrm{SE}[b_j]} \tag{10}$$

against the critical bounds of a standard normal distribution (e.g., ±1.96 for $\alpha = .05$, two-sided).[5] Values of z_j equal to or larger than the critical values lead to the rejection of $\mathrm{H}_0: \ \beta_j = 0$. Analogously, approximate 95% confidence intervals for the coefficients can be constructed with

$$b_j \pm 1.96\mathrm{SE}[b_j]. \tag{11}$$

Some alternative methods for making inferences about the coefficients in meta-regression models have been developed,[22] but we will not consider these methods further here.

3 | MODEL SELECTION

In practice, one often faces the problem that a large number of moderator variables have been measured, but it is unclear which of these moderators are actually related to the effect sizes. The problem of model selection in meta-regression can therefore be stated as follows. Let p denote the total number of moderators of interest and that could be included in the model. Now suppose m of these moderators (with $0 \le m \le p$) actually exert an influence on the effect sizes, while the remaining $p - m$ moderators do not. Let $T \subseteq \{1, ..., p\}$ denote the set with the m indices of the true moderators. Then the true model is given by

$$y_i = \beta_0 + \sum_{j \in T}\beta_j x_{ij} + u_i + \varepsilon_i \tag{12}$$

while the full model, containing "true" and "false" moderators, is given by

$$y_i = \beta_0 + \sum_{j \in T}\beta_j x_{ij} + \sum_{j \notin T}\beta_j x_{ij} + u_i + \varepsilon_i, \tag{13}$$

where, in truth, $\beta_j = 0$ for $j \notin T$. The goal is then to identify model (12) as the true model via some kind of model selection method.

We should clarify that model (12) is assumed to be the true model only in the sense that it reflects the best approximation to a presumably much more complicated reality given the available data. The actual data generating mechanism is likely to include moderators that have not actually been measured and/or coded and may not even be a linear model of the type described above. In addition, the moderators that have been coded are often just proxies or surrogates for the actual variables involved in the data generating mechanism. Therefore, whenever we refer to the true model and the true moderators, we do not mean to imply that the model or moderators reflect the true data generating mechanism, but that they are the best available approximation thereof given the moderators extracted from the studies and the hypothesized shape of their association with the observed effects.

Also, the particular application of meta-regression we are focused on here considers the p moderator variables as more or less equally relevant potential predictors of the effect sizes and the goal is to distinguish the true moderators of the effect sizes from the false ones. In other cases, one might be interested in testing a priori formulated hypotheses about specific moderators with a given meta-analytic dataset and/or one is interested in the relationship between a focal moderator (or a small number of them) and the effect sizes, while accounting for a set of other study characteristics that might confound the interpretation of the relationship between the effect sizes and the focal moderator(s). In such applications, model selection is not a primary concern and the methods to be described below are much less relevant.

3.1 | Model selection methods

We will consider four general approaches for identifying the set of true moderators and hence model (12) as the true model. The first approach includes simple methods that are commonly used in practice, namely univariate testing of the p moderator variables or testing of the p moderators in the context of the full model. The second approach examines the complete set of 2^p models that can be fitted given the p moderator variables and then uses various information criteria for model selection.[9,11] Next, we consider methods based on model averaging and multimodel inference over the set of 2^p models to identify the true moderators.[9,11] Finally, based on the information criteria, one can compute a "relative variable importance" for each of the moderators of interest, which can then be used for model selection. The various methods are described in more detail below.

3.2 | Selection via testing

In practice, relatively simple methods are commonly used for model selection in the meta-regression context. We will consider two such approaches based on Wald-type tests of the model coefficients. In the first approach, one simply fits p univariate meta-regression models (i.e., including one moderator variable at a time) and obtains the corresponding Wald-type tests of the coefficients. Based on the significance of these p tests, moderators are classified as either true or false. In the second approach, the full model including all p moderators is fitted. Again, Wald-type tests (now in the context of the full model) are used to categorize the moderator variables as being related to the effect sizes or not. With either testing approach, we can consider the true model as identified if we reject $\mathrm{H}_0: \beta_j = 0$ for all $j \in T$ and fail to reject $\mathrm{H}_0: \beta_j = 0$ for all $j \notin T$.

We do not consider the possibility of applying corrections for multiple testing (e.g., the Bonferroni correction), as this is not common practice in the meta-analytic context.[8] Moreover, while one could also select moderators based on the magnitude of their relationship with the effect size estimates (as reflected by the respective model coefficients) instead of their statistical significance, this is also not common practice and we will not consider this approach in the remainder of this article (it would also be difficult to formally define how exactly moderators should then be selected if one had to make decisions about their status as true or false moderators).

Finally, a general issue that one will often encounter in practice when conducting meta-regression analyses is missing information about certain moderator variables of interest for at least some of the studies. Missing information can lead to severe reductions in the "usable" data for fitting models with multiple moderator variables and/or lead to models that are based on different subsets of the dataset, which complicates the comparison of results across models. In the remainder of this manuscript, we will assume that the data are complete or at least that the studies with incomplete information have been removed before applying the methods discussed. We will come back to this issue in the discussion section.

3.3 | Selection via information criteria

Based on the p moderator variables, a total of $R = 2^p$ models can be fitted to the given data. While the true model should, on average, provide the best fit, we cannot use the log likelihoods directly for model selection, as the likelihood always increases as more moderators are added to the model. On the other hand, information criteria, which penalize the maximized likelihoods for model complexity, can be used for this purpose. The "best fitting" model (in the sense of making a trade-off between model fit and model complexity) is the one that minimizes a particular information criterion. From an information-theoretic perspective, model selection based on information criteria can also be described as a way for minimizing the loss of information (in a Kullback–Leibler sense) when approximating full reality by a fitted model.[9,11]

The most commonly known criterion for this purpose is the Akaike Information Criterion,[23] which is given by

$$\mathrm{AIC} = -2ll + 2(s+2), \tag{14}$$

where ll is either ll_{ML} or ll_{REML}, depending on the estimation method used, and s $(0 \leq s \leq p)$ denotes the number of moderators included in the model (hence, the model contains $s+2$ parameters, counting the model intercept and τ^2).

Another commonly used criterion is the Bayesian Information Criterion,[24] which is given by

$$\mathrm{BIC} = -2ll + (s+2)\ln\left(k^*\right), \tag{15}$$

where $k^* = k$ for ML estimation and $k^* = k - s - 1$ for REML estimation. When $k^* \geq 8$, the BIC penalizes the model fit more heavily than the AIC and therefore should tend to select models with fewer fixed effects.

Finally, we will consider a third criterion, a finite sample size (second-order bias) corrected version of the AIC,[25,26] given by

$$\mathrm{AICc} = -2ll + 2(s+2)\left(\frac{k^*}{k^* - (s+2) - 1}\right), \tag{16}$$

where $k^* = \max(k,\ s+4)$ for ML estimation and $k^* = \max(k - s - 1,\ s+4)$ for REML estimation (these definitions of k^* ensure that the additional multiplicative factor in the AICc is always ≥ 1). As $k/(s+2)$ increases (i.e., the ratio of sample size to the number of parameters), AICc converges to the AIC from above. However, in situations where k is small relative to the number of parameters, AICc will again tend to favor models with fewer fixed effects when compared to the AIC.

Regardless of the criterion, we then simply consider the true model as identified if it corresponds to the best model, that is, the model with the smallest value for the chosen information criterion. However, several issues are of note here when using REML estimation for model selection. First, the restricted log likelihood function (8) is obtained by taking linear combinations (contrasts) of the elements in $\boldsymbol{y}$, such that the transformed data are free of the fixed effects in $\boldsymbol{\beta}$.[27,28] Since the appropriate transformation depends on $\boldsymbol{X}$, models with different fixed effects will require different transformations, leading to restricted likelihoods that are technically not directly comparable. Consequently, likelihood ratio tests based on (8) are not appropriate for comparing models with different fixed effects.[13,14] However, some recent simulations suggest that model selection based on information criteria computed with (8) may still be a valid strategy.[15]

Second, while the restricted log likelihood function technically only contains one unknown parameter (i.e., τ^2), we do count the regression coefficients for the moderators as additional parameters when computing the information criteria under REML estimation, as otherwise we would always select the full model (including all p moderators) as the optimal one.

Finally, note that the second term (i.e., $\frac{1}{2}\ln|\boldsymbol{X}'\boldsymbol{X}|$) in the restricted log likelihood function (8) does not depend on τ^2 (or $\boldsymbol{\beta}$) and hence is irrelevant for maximizing the restricted log likelihood. Consequently, this term is often omitted by software when computing and reporting ll_{REML}. However, when using REML estimation for model comparisons involving different sets of moderators, the relevance of including this term in the computation of ll_{REML} is unclear.[15] Therefore, we define the maximized log likelihood with and without the second term in (8) as ll_{REMLf} and ll_{REMLr}, respectively, and can compute the information criteria with respect to both values. In total then, selection via information criteria can be conducted in nine different ways, by computing one of the three different information criteria (i.e., AIC, BIC, AICc) based on one of the three different likelihood functions (i.e., ll_{ML}, ll_{REMLf}, ll_{REMLr}).

3.4 | Selection via multimodel inference

Approaches that simply test coefficients (either univariately or in the context of the full model) do not take model uncertainty into consideration. As an alternative, we can base our inferences on all available models, using model averaging to obtain parameter estimates (and corresponding standard errors) that properly reflect this uncertainty. This multimodel inference approach works as follows.

First, based on a particular information criterion (e.g., AIC with ML estimation), we estimate the probability that each of the $R = 2^p$ models in the candidate set is the best model (in a Kullback–Leibler sense) with

$$w_r = \frac{\exp\left(-\frac{1}{2}\Delta_r\right)}{\sum_{r=1}^{R}\exp\left(-\frac{1}{2}\Delta_r\right)}, \tag{17}$$

where $\Delta_r = \mathrm{IC}_r - \mathrm{IC}_{\min}$, IC_r denotes the information criterion for the rth model, and $\mathrm{IC}_{\min}$ the value of the information criterion for the model with the smallest value (hence $\Delta_r = 0$ for the model selected as the best model according to the model selection strategies described in the previous section). These model probabilities are also commonly referred to as Akaike weights and, given particular priors, reflect posterior model probabilities in a Bayesian framework.[9,11] Note that these "Akaike weights" can be computed with any one of the information criteria discussed earlier (and based on the three different likelihood functions) and hence the name does not exclusively refer to the use of the AIC for their computation.

Instead of simply stopping here (and declaring moderators as true versus false depending on whether they are part of the model with $\Delta_r = 0$), we now proceed to the second step in multimodel inference. Here, model averaged parameter estimates for each coefficient are obtained with

$$\bar{b}_j = \sum_{r=1}^{R} w_r b_{rj}, \tag{18}$$

where b_{rj} denotes the estimated value of β_j in the rth model. Therefore, the estimated model coefficient for the jth moderator variable is then based on the entire collection of R models, with weights assigned in accordance to the estimated probabilities that each model is in fact the best model.

When computing $\bar{b}_j$, a decision needs to be made how to handle models that actually do not contain the jth moderator as one of the predictor variables (when considering all 2^p models, only half of the models will actually contain the jth moderator). Two strategies are commonly used to handle this[9,29]: When computing $\bar{b}_j$, we can only consider the subset of models that actually estimated β_j (doing so requires renormalizing the model weights, so that they sum to 1 for a given subset) or alternatively we can set the coefficient equal to zero for models that do not actually estimate β_j. The latter approach can be motivated on two grounds. First, the omission of a particular predictor from a model is in essence equivalent to assigning a value of zero to the corresponding coefficient. Second, setting the coefficient to zero in models that omit the jth moderator variable results in shrinkage of the model averaged parameter estimate that may counteract model selection bias.[11] We therefore only consider the second approach, but return to this issue in the discussion section.

An estimate of the variance of the model averaged parameter estimate accounting for both sampling error and model uncertainty can then be obtained with

$$\mathrm{Var}\left[\bar{b}_j\right] = \sum_{r=1}^{R} w_r \left(\mathrm{Var}\left[b_{rj}\right] + \left(b_{rj} - \bar{b}_j\right)^2 \right), \tag{19}$$

where $\mathrm{Var}[b_{rj}]$ denotes the estimated sampling variance of the jth model coefficient from the rth model.[11] When the rth model does not actually contain the jth model coefficient, then $\mathrm{Var}[b_{rj}] = 0$ (and $b_{rj} = 0$). The square-root of (19) provides an estimate of the standard error of the model averaged parameter estimate (i.e., $\mathrm{SE}\left[\bar{b}_j\right]$).

Finally, we can draw inferences about the relevance of each model coefficient, using (18) and (19) to construct the test statistic

$$\bar{z}_j = \frac{\bar{b}_j}{\mathrm{SE}\left[\bar{b}_j\right]}, \tag{20}$$

which we compare against the critical bounds of a standard normal distribution (e.g., ±1.96 for $\alpha = .05$, two-sided). Values of $\bar{z}_j$ equal to or larger than the critical values are again taken as evidence that the jth moderator variable is related to the effect sizes and is therefore classified as a true moderator. As with selection via information criteria, the multimodel inference approach can be conducted in nine different ways, depending on the information criterion and likelihood function used.

3.5 | Selection via relative variable importance

Instead of testing, we can also use the Akaike weight of each model to calculate the relative importance of each moderator, that is, the relative variable importance (RVI) across the candidate set of models, which for a particular moderator is the sum of the Akaike weights of all models in which that moderator occurs. Formally, letting $I_{rj} = 1$ if the jth moderator is included in the rth model and 0 otherwise,

$$\mathrm{RVI}_j = \sum_{r=1}^{R} w_r I_{rj} \tag{21}$$

is the sum of the Akaike weights for the jth moderator for all models in which the moderator appears. In practice, a sufficiently high RVI (often 0.5 or 0.8) is taken as evidence that a particular moderator is valuable for inference, although we are not aware of strong theoretical support for using any particular RVI criterion. Here, we explore both 0.5 and 0.8 as criteria, in each case taking an RVI greater than or equal to those values as evidence that a moderator is related to the effect sizes. As in testing, if $\text{RVI}_j \geq 0.5$ (or 0.8) for all $j \in T$ and $\text{RVI}_j < 0.5$ (or 0.8) for all $j \notin T$, we consider the true model as identified. The two RVI criteria can be crossed with the nine different combinations of information criteria and likelihood functions, leading to 18 different approaches that can be used for selecting moderators based on their RVI values.

3.6 | Model coefficient estimation

The goal of the methods described above is to distinguish the true from the false moderators, that is, we would like to identify model (12) as the true model. Even if these methods correctly select the true moderators (and none of the false ones), these procedures do not automatically provide estimates of the model coefficients in this model. For example, after univariate testing of each moderator, suppose that some are found to be significant (and hence selected as true moderators) and others are not. What is then an appropriate estimate of the strength and direction of the relationship between a selected moderator and the effect sizes? One possibility is to use the model coefficient from the univariate model including this moderator as the estimate. Alternatively, one could consider fitting an additional model that includes all of the selected moderators and using the coefficients from this model as the estimates of the respective relationships. Similarly, when using full model testing, should we report the results from the full model (including both the supposedly true and false moderators) or should we refit the model only including the true ones?

Similar issues arise when selecting moderators based on multimodel inference or their relative variable importance. The latter approach does not involve the size of the model coefficients at all, while the former makes use of coefficients that are weighted averages across all 2^p models. Should one report these averages as the best estimates of the relationships or should one refit a model that includes only the selected moderators? The only approach that inherently avoids these questions is selection via the information criteria, as it selects the true moderators based on their appearance in the "best" model (i.e., the one with the lowest value of the chosen information criterion) and it then seems quite natural to also report the coefficients from this model as the estimates of the relationship between the moderators and effect sizes. However, our goal in the present paper is not to provide answers to the questions raised above, but to evaluate the ability of the various model selection methods for distinguishing the true from the false moderators. Once we have examined the performance of the methods for this purpose, we can return to the question of how to estimate the relationship between the true moderators and the effect sizes in the discussion section.

4 | EXAMPLE

As an illustration, we applied the various approaches described above to a dataset of 80 effect size estimates obtained from studies examining the influence on plant biomass of inoculation with mycorrhizal fungi, which are soil-inhabiting root symbionts of plants that have the potential to improve plant growth. The dataset and R code to reproduce the following analyses are available at the Open Science Framework (https://osf.io/3d8u5/). Note that the dataset is used for illustration purposes only and is actually part of a much larger database,[30] including over 4000 estimates. Here, we focus on a subset of the data containing only estimates for corn, that is, the plant *Zea mays*, inoculated with one of two different genera of arbuscular mycorrhizal (AM) fungi, either *Funneliformis* or *Rhizophagus*. The subset was selected to avoid some of the additional complexities in the actual data structure (e.g., many plant species and fungi, phylogenetic relatedness, nesting and non-independence due to multiple estimates obtained from the same study and/or a shared control condition). The outcome measure used for the meta-analysis is the log response ratio,[31] with positive values reflecting an increase in mean biomass in plants receiving the mycorrhizal inoculation compared to non-inoculated plants.

A random-effects model fitted to the data using REML estimation yields an aggregated effect size estimate of $\hat{\mu} = 0.59$ (with 95% CI: 0.37–0.80). After exponentiation, we therefore estimate that mycorrhizal inoculation increases plant biomass on average by a factor of 1.80 (i.e., 80%; 95% CI: 1.45–2.23). However, there appears to be considerable heterogeneity in the effect sizes, with τ^2 being estimated at 0.523 in this model.

A large number of variables may systematically influence plant response to inoculation, but we focus here on four potential moderators. Besides the mycorrhizal fungus (FUN) used (k for *Funneliformis*: 29; k for *Rhizophagus*: 51), the dataset includes information about whether phosphorus fertilizer was added (FP; k for no: 54; k for yes: 26), whether nitrogen fertilizer was added (FN; k for no: 22; k for yes: 58), and whether or not the background soil was sterilized prior to mycorrhizal inoculation (STER; k for no: 16; k for yes: 64). Due to the observational nature of these data, the four factors were only partially crossed, with some combinations being much more prevalent than others (and some not represented at all). Computing correlations (phi coefficients) after dummy-coding each factor yielded values in the range of −0.14 to 0.53. See Table 1 for the full correlation matrix.

Table 2 shows the values of the AICc and model probabilities (i.e., w_r) for the $2^4 = 16$ possible models (considering only main effects and no interactions) when using REML estimation for the model fitting and computing the maximized restricted log-likelihood with ll_{REMLr}. The model with the lowest AICc includes moderators FN and FP, with an associated model probability of 0.47. Therefore, based on model selection via information criteria, we would consider moderators FN and FP as relevant.

Table 3 shows the results when testing each moderator in a series of univariate models, when testing each moderator in the context of the full model (i.e., with all four predictors included simultaneously), and when testing each moderator using multimodel inference using model averaged parameter estimates and corresponding standard errors along with their relative variance importance values based on all 16 models. From Table 2, we can see how the coefficient and standard error of the FP moderator is obtained for each of these approaches (see model 8 for the univariate model approach, model 6 when testing moderators in the context of the full model, and the value in the last row for the multimodel inference approach). Univariate testing leads to the conclusion that all four moderators are related to the effect sizes. When testing the moderators in the context of the full model and when using the multimodel inference approach, we come to a very different conclusion, only finding support for the relevance of FP as a potential moderator.

Clearly, there is substantial support for the importance of FP, with the predictor appearing in the top eight models as ranked by the AICc (the model probabilities for these eight models add up to 0.99, which is therefore also the relative variable importance for this moderator). Similarly, all testing methods identify this variable as a relevant predictor. Support for the relevance of the other moderators is weaker, although FN is part of the top three models as ranked by the AICc, with a relative variable importance of 0.85. Also, univariate testing (just barely) finds this moderator to be significant. The remaining moderators, FUN and STER, are not found to be significant by full model testing or multimodel inference. Furthermore, their relative variance importance are 0.31 and 0.32, respectively, both of which suggest that these moderators are unrelated to the effect sizes.

TABLE 1 Correlation matrix (phi coefficients) of the four dichotomous moderator variables

	FUN	FP	FN	STER
FUN	1	−0.14	0.53	0.27
FP	−0.14	1	0.31	0.01
FN	0.53	0.31	1	0.53
STER	0.27	0.01	0.53	1

The example illustrates that conclusions regarding the relevance of particular moderators can depend on the method used for model selection. To determine whether a particular approach is preferable (in terms of being more likely to select the true model), we conducted a simulation study.

5 | SIMULATION STUDY

Given the choice among the different approaches (and the resulting potential for conflicting conclusions), we conducted a Monte Carlo simulation study to compare the accuracy of the various methods for identifying the true model under ML and REML estimation. We also examined the relevance of using ll_{REMLf} versus ll_{REMLr} for computing the restricted log likelihood when using the various information criteria for model selection. In the present section, we describe the methods used for the simulation study and the results obtained.

5.1 | Methods

Assume that each study included in the meta-analysis compared two experimental groups with respect to a quantitative dependent variable. We will use the mean difference as the effect size measure for the meta-analysis, which is given by $y_i = \bar{x}_{i1} - \bar{x}_{i2}$, where $\bar{x}_{i1}$ and $\bar{x}_{i2}$ denote the observed means of the dependent variable in the first (e.g., treated) and second (e.g., non-treated) group. Then $y_i \sim N\left(\theta_i, \sigma_i^2(1/n_{i1} + 1/n_{i2})\right)$, where $\theta_i = \mu_{i1} - \mu_{i2}$ is the true effect in the ith study, μ_{i1} and μ_{i2} denote the true means of the two experimental groups, n_{i1} and n_{i2} the group sizes, and σ_i^2 the true variance of the dependent variable in an individual study. Without loss of generality, we set $\mu_{i2} = 0$, $\sigma_i^2 = 1$, and assume $n_{i1} = n_{i2} \equiv n_i$.

We purposefully chose to use an effect size measure for the simulation study that fulfills all of the assumptions of the random- and mixed-effects models (2) and (3) instead of simulating other measures commonly used in meta-analyses, such as standardized mean differences, log response ratios, risk differences, log risk/odds ratios, and raw or r-to-z transformed correlation coefficients. Although these measures will have normal sampling distributions with known sampling variances asymptotically, these assumptions can break down in smaller samples and/or under particular circumstances (e.g., for

Model	Moderator(s)	AICc	w_r	b_{FP}	$SE\,[b_{\text{FP}}]$
1	FN + FP	203.210	0.47	−1.00	0.207
2	STER + FN + FP	205.161	0.18	−0.97	0.209
3	FUN + FN + FP	205.507	0.15	−0.92	0.224
4	FUN + STER + FP	207.540	0.05	−0.71	0.206
5	FUN + FP	207.566	0.05	−0.66	0.208
6	FUN + STER + FN + FP	207.611	0.05	−0.90[a]	0.226
7	STER + FP	208.534	0.03	−0.75	0.209
8	FP	212.524	0.00	−0.71[b]	0.220
9	FUN	213.047	0.00	0.00	0.000
10	FUN + STER	214.157	0.00	0.00	0.000
11	FUN + FN	215.739	0.00	0.00	0.000
12	STER	216.281	0.00	0.00	0.000
13	FUN + STER + FN	216.765	0.00	0.00	0.000
14	FN	217.803	0.00	0.00	0.000
15	STER + FN	218.329	0.00	0.00	0.000
16	–	218.485	0.00	0.00	0.000
Avg				−0.93[c]	0.248

[a] Estimated model coefficient for moderator FP from the model including all moderators simultaneously.

[b] Estimated coefficient for moderator FP from the univariate model.

[c] Model-averaged parameter estimate (and corresponding standard error) for moderator FP.

TABLE 2 Value of the AICc (based on ll_{REMLr}) for the 16 models fitted to the data examining the influence of mycorrhizal inoculation on plant biomass

TABLE 3 Results for testing each moderator in a series of univariate models, in the full model, when using multimodel inference, and their relative variance importance (RVI) values based on all 16 models

	Univariate testing			Full model testing			Multimodel inference			
	b	SE	z	b	SE	z	b	SE	z	RVI
FUN	0.68	0.222	**3.07**	0.23	0.270	0.84	0.11	0.236	0.47	0.31
FP	−0.71	0.220	**−3.21**	−0.90	0.226	**−3.96**	−0.93	0.248	**−3.73**	**0.99**
FN	0.49	0.248	**1.99**	0.57	0.342	1.66	0.68	0.388	1.76	**0.85**
STER	0.64	0.277	**2.30**	0.28	0.298	0.93	0.12	0.248	0.47	0.32

Note: Test statistics that exceed ±1.96 and RVI values that exceed 0.5 are shown in bold.

rare events), at which point specialized methods/models may need to be applied.[32-34] With our chosen measure, we can study the performance of the various model selection methods unencumbered by these issues. Moreover, if conditions are such that the assumptions underlying the models are at least approximately fulfilled for other measures, then there is no particular reason to believe that the results obtained below would not be at least roughly applicable to these other measures as well.

In each iteration of the simulation, we first simulated the sample sizes of the individual studies by drawing k values from $\chi^2_{\text{df}=4}$, a Chi-square distribution with 4 degrees of freedom, and then letting $n_i = \bar{n}\left(\left(\chi^2_{\text{df}=4} + 6\right)/20\right)$ (rounded to the nearest integer), so that $\bar{n}$ denotes the average (total) sample size of each study included in the meta-analysis. The sample size distributions generated in this manner have a minimum of $\bar{n}(3/10)$, a mean of $\bar{n}/2$, and are right-skewed as often encountered in practice.[35,36]

Next, we simulated the values of the moderator variables by drawing k sets of p values from a multivariate standard normal distribution, where all p variables (i.e., the true and false moderators) were correlated with each other with correlation equal to ρ. When $\rho = 0$, the simulated moderator variables were uncorrelated, while values of $\rho \neq 0$ correspond to the more commonly encountered situation where moderator variables are correlated with each other.

Finally, given the sample sizes and the moderator values, we then simulated the k observed effect sizes from

$y_i \sim N\left(\sum_{j=1}^{m}\beta_j x_{ij}, \tau^2 + 2/n_i\right)$ for a given value of τ^2, so that moderators $j = \{1, ..., m\}$ are the true and $j = \{m+1, ..., p\}$ are the false moderators. We set $\beta_j = \beta$ for $j = 1$ to m. Therefore, each of the m true moderators exerted the same amount of influence on the effect sizes. On the other hand, the remaining $p - m$ false moderators did not influence the effect sizes.

Within each iteration, we therefore obtained k values of y_i, the corresponding sampling variances $v_i = 2/n_i$, and the values of the p moderator variables corresponding to each of the k studies. The various model selection approaches described earlier were then used in an attempt to identify the true model. Therefore, we fitted each of the 2^p possible models to the data using ML and REML estimation and recorded the corresponding AIC, BIC, and AICc values. For REML estimation, both ll_{REMLf} and ll_{REMLr} were used for computing the restricted log likelihood and the resulting information criteria. We then recorded whether the true model was correctly selected according to the minimum AIC, BIC, and AICc criteria.

For the testing approaches, we recorded whether the set of significant and non-significant moderators corresponded to the set of true and false moderators in the true model, respectively. If all of the true moderators and none of the false moderators were significant, then this was considered a correct identification of the true model. The same approach was used when using the RVI of the variables as the selection method (using either 0.5 or 0.8 as the cutoff for considering a variable as important).

The following factors were examined in the simulation study: k (20, 30, 40, 60, 80), $\bar{n}$ (15, 30, 60, 120), ρ (0, 0.3, 0.6), β (0, 0.1, 0.2, 0.3, 0.4), τ^2 (0, 0.1^2, 0.2^2, 0.3^2, 0.4^2), p (4, 6, 8, 10) and m (either m was set equal to $p/2$ or held constant at $m = 2$). All factors were fully crossed, so that a total of $5 \times 4 \times 3 \times 5 \times 5 \times 4 \times 2 = 12{,}000$ conditions were examined (note that when $\beta = 0$, then the true model is actually the "empty model" that does not include any moderators, regardless of the value of m). For each condition, we simulated 1000 meta-analyses and then estimated the probability of identifying the true model by computing the proportion of iterations in which the true model was correctly identified by each of the methods described above.

We fully acknowledge that in real data, moderators are going to exhibit different degrees of correlation among each other (instead of assuming a constant value of ρ for each pair), will be a mixture of different variable types (e.g., integer, [semi-]continuous, categorical with two or more levels) with various types of distributions (instead of following a multivariate normal distribution), and the magnitude of the relationship between the moderators and the effect sizes will vary across moderators (instead of being constant for all true moderators). The simplifications above were made to keep the number of conditions manageable, which are already quite large. Moreover, while we could introduce variability into the value of ρ for pairs of moderators, simulate moderators from different distributions, and allow β_j to differ across true moderators, we would not expect this to have a substantial impact on the *relative* performance of the various methods.

For the design of the simulation study above, it can be shown that the total amount of heterogeneity in the true effects is equal to $\text{Var}[\theta_i] = m(1+(m-1)\rho)\beta^2 + \tau^2$. Furthermore, note that, on average, $n_i = \bar{n}/2$, so that, on average, $\bar{v} = 4/\bar{n}$. Based on these values, we can compute $I^2 = \text{Var}[\theta_i]/(\text{Var}[\theta_i] + \bar{v})$, that is, how much of the total amount of variability in the effect size estimates can, on average, be attributed to heterogeneity in the true effects. Therefore, when $\beta = 0$, the values of I^2 corresponding to the $\bar{n}$ and τ^2 values chosen for the simulation study range from 0 to 0.38 (with a median of 0.13) for $\bar{n} = 15$ and from 0 to 0.83 (with a median of 0.55) for $\bar{n} = 120$. When $\beta > 0$, the resulting I^2 values increase in accordance with β, m, ρ, and τ^2. Values of I^2 around 0.25, 0.50, and 0.75 are typically considered to reflect low, moderate, and high amounts of heterogeneity,[37] but when multiple strong moderators are exerting an influence on the effect sizes, values above 0.90 are not uncommon. Therefore, the values of the factors above were chosen in accordance with these considerations.

Furthermore, the simulation design implies that the amount of heterogeneity that can be accounted for by any individual true moderator is equal to $R^2 = \beta^2/(m(1+(m-1)\rho)\beta^2 + \tau^2)$. Therefore, as m, ρ, and τ^2 increase, the explanatory power of each individual moderator decreases. However, the total amount of heterogeneity that can be accounted for by all true moderators increases with m and ρ and is simply equal to 1 when $\tau^2 = 0$.

The simulation study was programmed in R.[38] Due to the large number of conditions and the highly intensive nature of the computations (especially when fitting all 2^p models), the simulation study was programmed to make use of multicore processing using parallelization. The simulation was run on a cluster computer using 126 cores simultaneously for the computations. Even making use of such computing power, completion time for the entire set of conditions was a bit over 5 days (roughly 16,000 core hours in total).

5.2 | Results

Given the large number of conditions and selection methods examined in the simulation, summarizing the findings is challenging. To obtain a broad overview, we

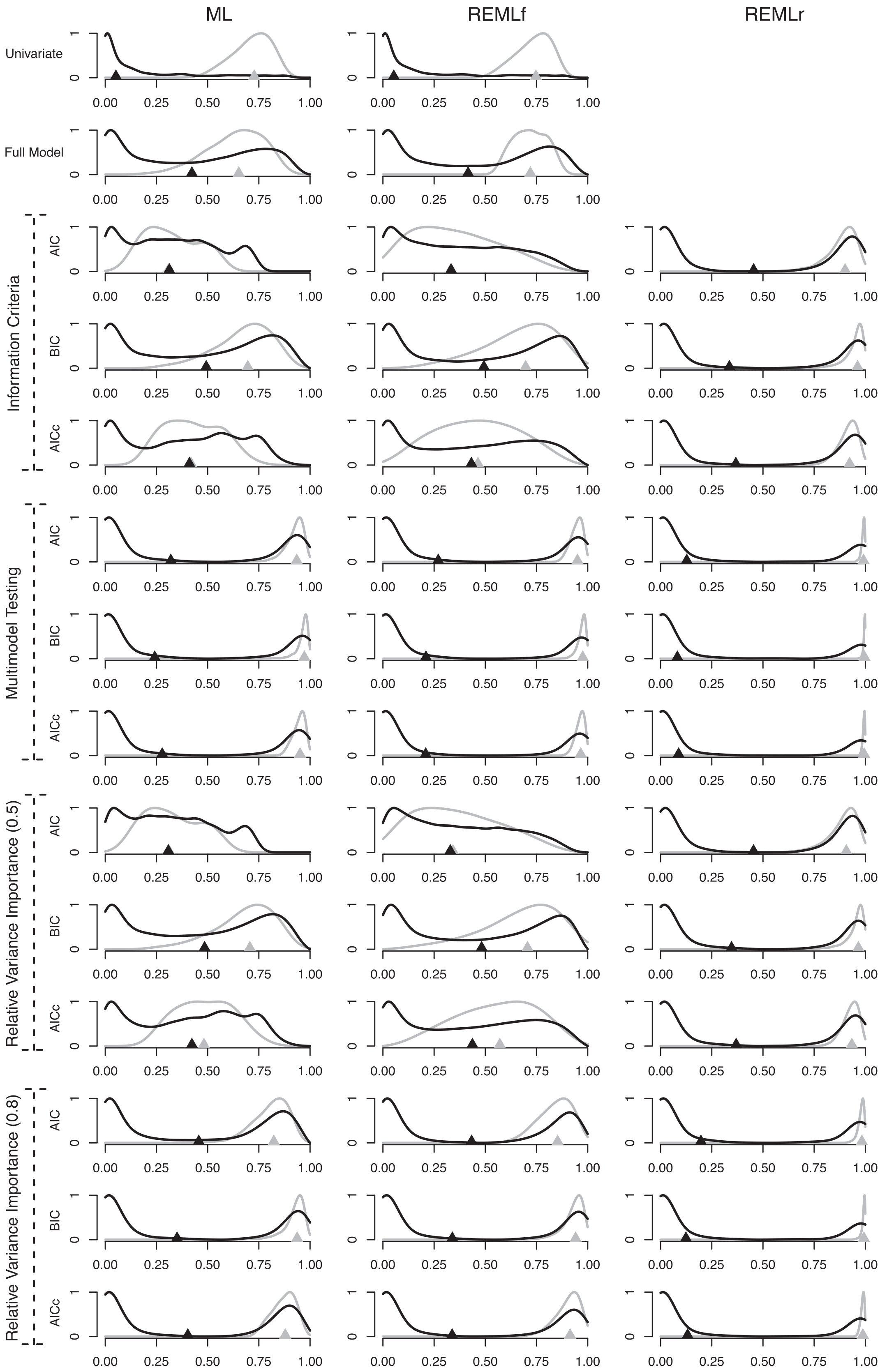

FIGURE 1 Kernel density plots of the probabilities of identifying the true model across all 12,000 conditions for each of the methods. Gray lines represent conditions where the true model was the empty model (i.e., $\beta = 0$), whereas black lines represent conditions where there was a non-zero association between the moderators and the effect sizes (i.e., $\beta > 0$). The triangles (gray triangle and black triangle) indicate the corresponding median probabilities

start by presenting kernel density plots of the probabilities of identifying the true model across all conditions for each of the methods in Figure 1 (40 different methods in total). The two lines in each plot distinguish the 2,400 conditions where the true model was the empty model (gray lines) from the 9,600 conditions where there was a non-zero association between the moderators and the effect sizes (black lines). The gray and black triangles in the plots correspond to the respective median probabilities. To make comparisons between these two classes of conditions and across methods easier, the densities were rescaled to have a maximum of 1.

Selection via univariate or full model testing showed adequate performance when the true model was the empty model (no plots are shown for ll_{REMLr} because these methods are not affected by the way the REML likelihood is computed). In many such conditions, these methods correctly identified the empty model in at least 50% of the cases and typically even with higher probabilities (the median probabilities were close to 75%). On the other hand, when there was an association between the moderators and the effect sizes, univariate testing had very low chances of identifying the true model in most conditions. Full model testing showed more promising performance. Here, the density was bimodal and appears to be composed of two "sub-densities," the first peaking just above 0%, which corresponds to conditions where identification of the true model is particularly difficult (i.e., various combinations of low k, low $\bar{n}$, high ρ, low β, high τ^2, and high p). The other part of the density peaks around 80%, but is quite left-skewed.

Several general conclusion can be reached about the methods based on information criteria. First, their performance was quite similar when computed based on ll_{ML} or ll_{REMLf}. Second, using the BIC was better at correctly identifying the empty model compared to the AIC, with the AICc typically falling in-between these two criteria at least in terms of the median performance. Moreover, all multimodel inference methods and all methods based on ll_{REMLr} performed exceptionally well in identifying the true model when none of the moderators were related to the effect sizes. Third, when the true model was not the empty model, performance of the various methods varied quite a bit, although we also typically see the bimodal shape in the densities as described above. In fact, for some methods, the bimodality was even more pronounced. In other words, depending on the condition, there was then either a very low or a very high probability of correctly identifying the set of moderators that were truly related to the effect sizes.

To more directly compare the model selection methods against each other, we determined the method that obtained the highest probability of identifying the true model for each condition. However, a particular method may turn out to be the best for a particular condition simply due to simulation error. Moreover, the best method may outperform other methods just by a small margin, which could be considered practically irrelevant. Therefore, for each condition, we determined the set of methods that were no more than five percentage points worse than the best method (e.g., if the best method had a 0.38 probability of identifying the true model in a particular condition, then all methods that had at least a 0.33 probability were among the set of best methods for this condition).1 We then computed the proportion of conditions in which each method was among the best method (again separately for conditions where $\beta = 0$ and where $\beta > 0$). The results are shown in Figure 2.

When the true model was the empty model, univariate and full model testing were never among the best methods in any of the conditions. On the other hand, some of the information criteria methods were quite often among the best methods, especially when computed based on the ll_{REMLr} function. In that case, using multimodel inference and the RVI with a cutoff value of 0.8 was always or almost always among the best performing methods for identifying the empty model, regardless of whether the AIC, BIC, or AICc was used. On the other hand, when selecting based on the minimum of the information criteria or an RVI with a cutoff value of 0.5, we again see that the BIC performed best, followed by the AICc and the AIC.

When the true model was not the empty model, univariate and full model testing were among the best methods for up to 20% of the conditions. However, the information-theoretic approaches always outperformed selection via univariate or full model testing, falling among the best methods for 26% to up to 49% of the conditions. While the specific information criterion and likelihood function played a relatively minor role in these results, selection via minimum information criteria or an RVI with a cutoff value of 0.5 tended to be among the best methods more often than selection via multimodel inference or when using an RVI with a cutoff value of 0.8.

To determine the factors that are most relevant for accounting for the differences between the various selection methods, we structured the results from the simulation study as a dataset with 12,000 (for conditions) × 40 (for methods) = 480,000 rows and then conducted a two-way analysis of variance (ANOVA) with the (arcsine square-root transformed) probabilities of identifying the true model as the outcome variable (this data set is also available at OSF, https://osf.io/3d8u5/). The selection method factor, all design factors of the simulation, and their interactions with the selection method factor were

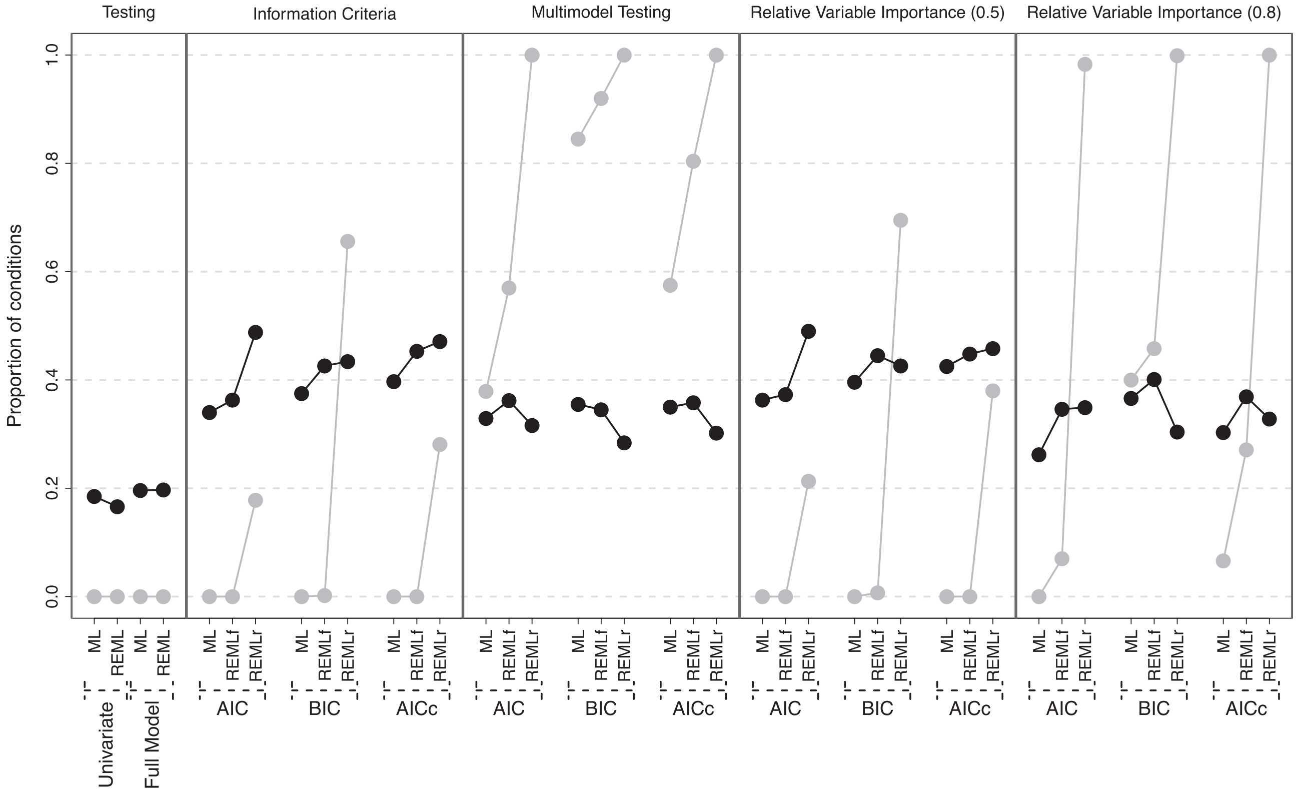

FIGURE 2 Proportion of conditions where each method was among the best methods (i.e., no worse than five percentage points than the best method). Gray lines represent conditions where the true model was the empty model (i.e., $\beta = 0$), whereas black lines represent conditions where there was a non-zero association between the moderators and the effect sizes (i.e., $\beta > 0$)

TABLE 4 η^2 values derived from the two-way ANOVA predicting the probabilities of selecting the true model based on the model selection method, all design factors, and their two-way interactions

Method	0.27							
k	0.10	0.47						
$\bar{n}$	0.07	0.01	0.42					
ρ	0.05	0.00	0.00	0.18				
β	0.51	0.23	0.17	0.10	0.85			
τ^2	0.02	0.00	0.07	0.00	0.06	0.30		
p	0.04	0.00	0.00	0.00	0.02	0.00	0.24	
m	0.01	0.01	0.00	0.00	0.02	0.00	0.02	0.05
	Method	k	$\bar{n}$	ρ	β	τ^2	p	m

Note: The diagonal and off-diagonal values display the η^2 values of the main effects and two-way interactions, respectively.

included as predictors in the model. Table 4 presents the η^2 values of the main effects (along the diagonal) and the two-way interactions (the off-diagonal elements). The most influential factor was the size of the coefficient of the moderators (i.e., β), which is also responsible for distinguishing conditions where the empty model was the true model from conditions where moderators were actually present, followed by the number of effect sizes (k) and the sample size of the primary studies ($\bar{n}$). Furthermore, the size of the coefficient showed strong interactions with the selection method, the number of effect sizes, and the sample size of the primary studies.

In Figure S1, we also provide plots of the probabilities of identifying the true model as a function of the various design factors for each of the 40 methods. Increases in k, $\bar{n}$, and β (for the non-empty model) were associated with higher probabilities, which is not surprising, since increases in these factors provide more/stronger evidence about the existence of relationships (or their absence). On the other hand, increases in ρ, τ^2, and p were associated with lower probabilities, which is also expected, since stronger correlations among the moderators, higher (residual) heterogeneity, and a larger number of potential moderators makes it more difficult to detect the true ones. Also, on average, probabilities tended to be slightly higher when m was fixed at 2 as opposed to conditions where m increased as a function of $p/2$.

Based on these findings from the ANOVA, we constructed Figure 3, which shows the performance of six representative methods as a function of the coefficient size (i.e., β) when k is equal to 20, 40, 60, and 80. We included selection via univariate and full model testing as reference benchmarks and used the AICc criterion combined with the ll_{REMLr} function for each of the information-theoretic approaches (Figures S2 and S3 provided as part of the Supporting Information are analogous figures using the AIC and BIC criteria, but show very similar patterns).

When the true model was the empty model (i.e., $\beta = 0$), we again see the superior performance of the information criteria methods compared to univariate or full model testing. While the latter two methods were able to identify the true model on average in about 70% of the cases, using multimodel inference and an RVI with a cutoff of 0.8 yielded essentially 100% correct identification rates regardless of k, closely followed by using the minimum AICc or an RVI with a cutoff of 0.5 for model selection.

When there is a weak association between the effect sizes and the moderators (i.e., $\beta = 0.1$), all methods yielded very low probabilities (i.e., ≤0.10), with only little improvements as k increased. Selection via full model testing may have a slight advantage over the remaining methods in this case especially when k is large. On the other hand, while larger coefficients and values of k led to increases in the probabilities for all methods (except for univariate testing, whose performance flattened out at around 20%), selection via the minimum AICc or an RVI with a cutoff of 0.5 outperformed full model testing especially when k and β were large. In fact, while all information criteria methods converged to 100% rates with increases in k and β, this does not appear to be the case for full model testing, which seems to reach a maximum identification rate around 80%. Figures S4–S7 provide analogous figures to Figure 3, but separately for each value of $\bar{n}$, showing that this maximum is also reached

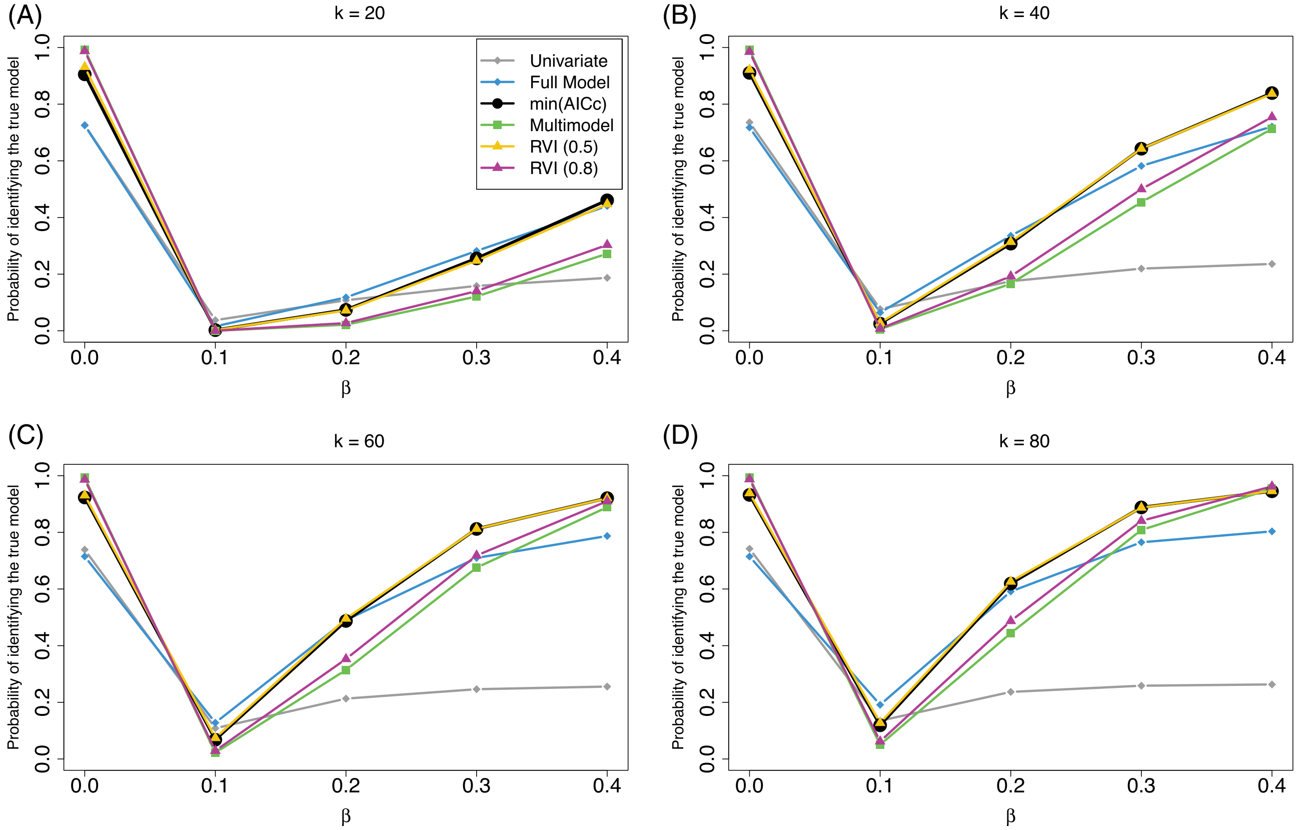

FIGURE 3 Probabilities of identifying the true model of six model selection methods as a function of β and k. The methods shown are selection via univariate and full model testing and the information-theoretic approaches using the AICc criterion combined with the ll_{REMLr} function. Each line represents the probabilities of a model selection method averaged over the remaining factors. (a) $k = 20$. (b) $k = 40$. (c) $k = 60$. (d) $k = 80$ [Colour figure can be viewed at wileyonlinelibrary.com]

even when the sample size of the primary studies increases.

6 | DISCUSSION

In this paper, we describe methods for model selection in meta-regression with particular emphasis on information-theoretic approaches. In doing so, we hope to introduce researchers unfamiliar with such methods to a new set of tools that may be useful in applied research. However, one should only consider applying novel methods if evidence suggests that there are tangible benefits to using these approaches over more conventional methods for model selection, such as testing moderator variables one at a time in a series of meta-regression models (i.e., what we denote "univariate testing") or testing moderator variables in a single meta-regression model (i.e., "full model testing"). We therefore also conducted a simulation study to examine the performance of the various model selection methods for identifying the set of moderators that are actually related to the effect sizes.

Across a wide variety of conditions, the results show that the information-theoretic approaches often outperform the more conventional methods. This was especially apparent for conditions where none of the moderators were truly related to the effect sizes (i.e., when the "empty model" was the true model), but could also be seen when there was a relationship between some of the moderators and the effect sizes. Only when the relationship was very weak was there a slight advantage to using a full model testing approach, although differences between methods were small and all methods had rather low probabilities of identifying the true model in such conditions.

Of all methods evaluated, univariate testing performed especially poorly even when the relationship between moderators and effect sizes was strong and the number of studies was large. This finding is not surprising, given that correlated moderator variables were simulated in many conditions. Univariate testing will then often find that "false" moderators are significant simply because they happen to be correlated (i.e., confounded) with the true moderators (a phenomenon also known as omitted-variable bias). At the same time, when testing a true moderator without other relevant moderators included in the model, at least part of the heterogeneity that could have been accounted for by the other moderators will then be subsumed into the random effects term used to model residual heterogeneity. This in turn will decrease the power to detect the true moderator. Hence, too many false and too few true moderators will be correctly identified when using univariate testing, leading to the poor performance of this model selection strategy. Although concerns about univariate testing of moderators have been raised before,[39] these concerns warrant repetition, given that univariate testing is still the dominant approach in meta-regression analyses.[8]

Full model testing can circumvent these problems, at least if we assume that all potentially relevant moderators are included in the model (or that those omitted are not correlated with the ones included in the model). In that case, the correlation among the moderators is correctly taken into consideration, avoiding omitted-variable bias. Consequently, this model strategy also fared much better in the simulation study, but the results also indicate that the probability of identifying the true model did not converge to 100% even when both the number of studies and the size of the regression coefficient increased.

There is a simple explanation for this finding. Even when the power to detect the true moderators is 100% so that all true moderators will be correctly identified, each false moderator tested incurs a certain chance of committing a Type I error. The more such false moderators are tested, the higher the probability that at least one of them will turn out to be significant. Given the design of the simulation study, there were either $f = 2$, 3, 4, or 5 false moderators in conditions where the number of true moderators was set to half of the number of moderators tested or there were either $f = 2$, 4, 6, or 8 false moderators in conditions where the number of true moderators was fixed at 2. The probability that at least one of the false moderators turns out significant is $1-(1-\alpha)^f$, which we can compute for each value of f above (with $\alpha = .05$). Averaging the resulting values yields approximately 0.19. Hence, we should see that the probability of correctly identifying the true model converges to $1-0.19 = 0.81$, which is exactly what Figure 3 suggests (see panel (d) for $k = 80$ and $\beta = 0.4$).

An obvious solution to this problem is to apply a multiple testing correction for each of the moderators tested in the context of the full model. Although this will also reduce power to detect true moderators, 100% power should be restored with sufficiently large k and β. There will still be a small probability that at least one false moderator is selected and hence a perfect model identification probability cannot be achieved, but the discrepancy should be minor. However, we did not explore this strategy in the simulation study, as the use of multiple testing corrections in the context of meta-regression analyses is rare.[8]

In contrast, all methods based on information criteria achieved perfect or near perfect model identification probabilities when k and β increased. This also applies to the multimodel testing approach, which seems counterintuitive given that we did not use a multiple testing correction in the context of this strategy either. However, this model selection strategy goes beyond full model testing by also using model-averaged parameter estimates and standard errors for doing so. The way the model-averaged parameter estimates are computed—by assigning zero to the coefficient for a particular moderator variable in models where the moderator does not appear—will automatically lead to a shrinkage effect when models without a particular moderator receive large Akaike weights.[11] This effect can also be seen in the illustrative example (cf. Table 3) especially for the coefficients corresponding to the mycorrhizal fungus (FUN) and soil sterilization (STER) moderators, given that almost 70% of the total weight (i.e., $(1 - \mathrm{RVI}) \times 100\%$) is placed on models in which these moderators do not appear. In this sense, multimodel testing shares some properties with other penalization techniques, such as ridge regression and the lasso,[40,41] although these methods have not been extended to the meta-regression context.

Simply selecting a model based on the minimum of a particular information criterion or using the RVI values with a certain cutoff for variable selection works equally well as the multimodel testing approach when k and β are large. For all these strategies, an increasingly large weight will then be placed on the model that corresponds to the true model, leading to perfect model identification probabilities. In less favorable circumstances, the best method seems to depend on whether one favors a method that has better chances of detecting that none of the moderators are actually related to the effect sizes (in which case multimodel testing or using the RVIs with a cutoff of 0.8 appear preferable) or whether one would like to optimize one's chances that the method will correctly sort out which moderators are and which ones are not related to the effect sizes (in which case selection based on the minimum of an information criterion or using the RVIs with a cutoff of 0.5 would be the better choice).

Cautiously, we would therefore suggest that if there is convincing a priori theoretical support for the potential influence of the moderators considered in the search, the latter two strategies should be preferred. Among these, the minimum information criterion approach has the advantage of being very easy to apply. Moreover, as discussed in Section 3.6, this approach automatically provides the estimates of the model coefficients for the selected moderators without any additional steps. However, compared to univariate or full model testing, this approach is computationally much more demanding, as it requires fitting 2^p models. When $p = 20$, it may take some hours to do so (unless one takes extra steps to optimize/parallelize the computations). For $p = 30$, this approach may become computationally prohibitive.

At the same time, we are fully aware of cautions in the literature against simply using all possible models as candidates for the true model instead of conducting an analysis based on a smaller set of candidate hypotheses and their corresponding models.[9,11] Even though we used this approach in the simulation study, this should be seen as an abstraction done for the purposes of simplifying the simulation study. Also, potential moderators should always be a priori specified[7] (ideally with a hypothesis why and how they might be related to the effect sizes) and hence the p moderators examined in the simulation study could be considered to be a selection based on a larger number of potential moderators.

A noteworthy finding of the simulation study is that model selection via information criteria computed based on the REML function appears to be a valid strategy. In principle, this finding goes against the common wisdom that REML functions (and hence information criteria) computed based on models with different fixed effects are not comparable.[13,14] The present findings are however in line with those of Gurka,[15] who also presented results to the contrary, but in a different modeling context. However, when using information criteria based on REML estimation, one has to carefully consider how exactly the restricted likelihood will be computed, that is, whether the $\frac{1}{2}\ln|\boldsymbol{X}'\boldsymbol{X}|$ term should be included in the likelihood function or not. While the findings from the simulation study are somewhat mixed and the right approach might depend on the strategy and information criterion used, the results clearly indicate that omission of the term is important to obtain high probabilities of correct model identification when the empty model is actually the true model.

Regardless of the chosen strategy, a practical issue that frequently arises in meta-regression analyses is missing data. In particular, while some moderator variables are easy to extract (e.g., the publication year of a study) and will be essentially complete, missing data on other moderator variables is a common occurrence. A comparison of models containing different subsets of moderator variables will then be hampered by the fact that the models will be based on different subsets of the data, which makes their likelihoods and hence the information criteria incomparable. On the other hand, using only the set of studies with complete information on all moderator variables of interest will typically lead to a substantial reduction in the number of studies included in the analysis. Note that this will also automatically happen when using full model testing, as "listwise deletion" is the default behavior in all software for conducting meta-regression analyses that we are aware of. In fact, we suspect that the prevalence of univariate testing may at least in part stem from the desire of authors to maximize the number of studies included in each meta-regression model. However, as the current results show, univariate testing cannot be recommended as a general model selection strategy. Alternatively, one could consider methods for imputing missing moderator values,[42,43] although very little work has examined the performance of such methods in the meta-analytic context so far (but see the findings from Ellington et al.[44]). How to combine imputation methods with model selection strategies including the ones discussed in the present paper could therefore be the subject of future research.

In our simulation study, we also did not consider more complex data structures one may often encountered in practice. In particular, in many meta-analyses, one can extract multiple effect size estimates from some or all of the studies, leading to dependencies in the data, which can be addressed for example by means of appropriate multilevel/multivariate models and/or cluster-robust inference methods.[45,46] While there are no fundamental difficulties in computing information criteria for more complex models, and hence the various model selection methods described can and have already been applied in this context as well,[47] we would welcome further research in this direction.

Furthermore, we should note that the "true model" was actually included in the set of candidate models in our simulation study, although this is unlikely to correspond to reality, which is more complex than any statistical model we might formulate.[9,11] Hence, in practice, the true model is not going to be part of the candidate set. However, this scenario is difficult to simulate and would require that we quantify to what extent each model approximates some true data generating mechanism. At the same time, by allowing for (residual) heterogeneity, random- and mixed-effects models in meta-analysis already include a term that is meant to capture any influences on the effect sizes that the model is unable to account for. Conditions where $\tau^2 > 0$ are therefore scenarios where even the true model from the candidate set could just be considered to be the best approximation to the actual data generating mechanism, however complex it may be.

In conclusion, the present article provides some initial evidence that conventional methods for model selection, such as univariate and full-model testing, may be outperformed by information-theoretic approaches. The latter are more often among the set of best methods across all of the conditions simulated and can have higher probabilities for identifying the true model under particular scenarios. We recommend that authors of meta-analyses involving meta-regression analyses consider the use of these methods, especially as an alternative to univariate testing. The methods can be easily implemented in R using the metafor package[48] in combination with the glmulti[49] or MuMIn[50] packages. The R code provided as part of the illustrative example (at https://osf.io/3d8u5/) can be easily adapted to other applications. Another fully worked example illustrating the use of these methods can be found at https://www.metafor-project.org/doku.php/tips:model_selection_with_glmulti_and_mumin.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

AUTHOR CONTRIBUTIONS

James Umbanhowar, Jason D. Hoeksema, and **Wolfgang Viechtbauer** conceived the initial idea. **James Umbanhowar** and **Wolfgang Viechtbauer** wrote the simulation code. **Ozan Cinar** carried out the simulation study, processed the results, and drafted the manuscript. All authors contributed to editing the manuscript; all read and approved the final manuscript.

ORCID

Ozan Cinar https://orcid.org/0000-0003-0329-1977
James Umbanhowar https://orcid.org/0000-0002-8251-9388
Jason D. Hoeksema https://orcid.org/0000-0003-2338-8442
Wolfgang Viechtbauer https://orcid.org/0000-0003-3463-4063

ENDNOTE

[1] The 0.05 margin is admittedly arbitrary, but slight variations in the margin did not materially affect the conclusions.

REFERENCES

1. Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd ed. New York: Russell Sage Foundation; 2009.
2. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *Br Med J*. 1994;309(6965):1351-1355.
3. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med*. 1995;14(4):395-411.
4. Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med*. 2002;21(4):589-624.
5. Raudenbush SW. Analyzing effect sizes: random-effects models. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd ed. New York: Russell Sage Foundation; 2009:295-315.
6. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med*. 1999;18(20):2693-2708.

7. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21(11): 1559-1573.
8. Tipton E, Pustejovsky JE, Ahmadi H. Current practices in meta-regression in psychology, education, and medicine. *Res Synth Methods*. 2019;10(2):180-194.
9. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer; 2002.
10. Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP. Why do we still use stepwise modelling in ecology and behaviour? *J Animal Ecol*. 2006;75(5):1182-1189.
11. Anderson DR. *Model Based Inference in the Life Sciences: A Primer on Evidence*. 2nd ed. New York: Springer; 2007.
12. Chatfield C. Model uncertainty, data mining and statistical inference. *J Roy Stat Soc Ser A*. 1995;158(3):419-466.
13. Pinheiro JC, Bates DM. *Mixed-Effects Models in S and S-PLUS*. New York: Springer; 2000.
14. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. New York: Springer; 2000.
15. Gurka MJ. Selecting the best linear mixed model under REML. *Am Stat*. 2006;60(1):19-26.
16. Shi P, Tsai CL. Regression model selection: a residual likelihood approach. *J Roy Stat Soc Ser B*. 2002;64(2):237-252.
17. Fernandez GC. Model selection in PROC MIXED: A user-friendly SAS macro application. Paper presented at the SAS Global Forum; 2007.
18. Borenstein M. Effect sizes for continuous data. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd ed. New York: Russell Sage Foundation; 2009:221-235.
19. Fleiss JL, Berlin JA. Effect sizes for dichotomous data. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd ed. New York: Russell Sage Foundation; 2009:237-253.
20. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J Educ Behav Stat*. 2005;30(3):261-293.
21. Nocedal J, Wright SJ. *Numerical optimization*. 2nd ed. New York: Springer; 2006.
22. Viechtbauer W, Lopez-Lopez JA, Sanchez-Meca J, Marin-Martinez F. A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychol Methods*. 2015;20(3):360-374.
23. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control*. 1974;19(6):716-723.
24. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461-464.
25. Hurvich CM, Tsai CL. Regression and time series model selection in small samples. *Biometrika*. 1989;76(2):297-307.
26. Hurvich CM, Tsai CL. Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika*. 1991;78(3):499-509.
27. Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc*. 1977;72(358):320-338.
28. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. *Biometrika*. 1971;58(3):545-554.
29. Lukacs PM, Burnham KP, Anderson DR. Model selection bias and Freedman's paradox. *Ann Inst Stat Math*. 2010;62(1):117-125.
30. Chaudhary VB, Rúa MA, Antoninka A, et al. MycoDB, a global database of plant response to mycorrhizal fungi. *Sci Data*. 2016;3:160028.
31. Hedges LV, Gurevitch J, Curtis PS. The meta-analysis of response ratios in experimental ecology. *Ecology*. 1999;80(4):1150-1156.
32. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 1959; 22(4):719-748.
33. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis*. 1985;27(5):335-371.
34. Stijnen T, Hamza TH, Ozdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Stat Med*. 2010;29(29):3046-3067.
35. Davey J, Turner RM, Clarke MJ, Higgins JP. Characteristics of meta-analyses and their component studies in the Cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC Med Res Methodol*. 2011;11:160.
36. Sanchez-Meca J, Marin-Martinez F. Testing continuous moderators in meta-analysis: a comparison of procedures. *Br J Math Stat Psychol*. 1998;51(2):311-326.
37. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *Br Med J*. 2003;327(7414):557-560.
38. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria; 2020.
39. Lipsey MW. Those confounded moderators in meta-analysis: good, bad, and ugly. *Ann Am Acad pol Soc Sci*. 2003;587:69-81.
40. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Dent Tech*. 1970;12(1):55-67.
41. Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Stat Soc Ser B*. 1996;58(1):267-288.
42. Pigott TD. Methods for handling missing data in research synthesis. In: Cooper H, Hedges LV, eds. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation; 1994:163-176.
43. Pigott TD. Missing predictors in models of effect size. *Eval Health Prof*. 2001;24(3):277-307.
44. Ellington EH, Bastille-Rousseau G, Austin C, et al. Using multiple imputation to estimate missing data in meta-regression. *Method Ecol Evol*. 2015;6(2):153-163.
45. Moeyaert M, Ugille M, Beretvas SN, Ferron J, Bunuan R, Noortgate V. Methods for dealing with multiple outcomes in meta-analysis: a comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *Int J Soc Res Methodol*. 2017;20(6):559-572.
46. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Stat Med*. 1998;17(22):2537-2550.
47. Hoeksema JD, Bever JD, Chakraborty S, et al. Evolutionary history of plant hosts and fungal symbionts predicts the strength of mycorrhizal mutualism. *Commun Biol*. 2018;1(1):116.
48. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48.
49. Calcagno V. glmulti: Model selection and multimodel inference made easy. R package version 1.0.8; 2020.

50. Bartoń K. MuMIn: Multi-model inference. R package version 1.43.17; 2020.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Cinar O, Umbanhowar J, Hoeksema JD, Viechtbauer W. Using information-theoretic approaches for model selection in meta-analysis. *Res Syn Meth*. 2021;12(4):537–556. https://doi.org/10.1002/jrsm.1489