Received: 10 May 2022 | Accepted: 14 December 2022

DOI: 10.1111/bmsp.12299

the british psychological society
promoting excellence in psychology

ARTICLE

# Heterogeneous heterogeneity by default: Testing categorical moderators in mixed-effects meta-analysis

Josue E. Rodriguez[1] | Donald R. Williams[1,2] | Paul-Christian Bürkner[3]

[1]University of California, Davis, California, USA

[2]NWEA, Portland, Oregon, USA

[3]Cluster of Excellence SimTech, University of Stuttgart, Stuttgart, Germany

**Correspondence**

Josue E. Rodriguez, University of California, Davis, Davis, CA, USA.
Email: jerrodriguez@ucdavis.edu

**Funding information**

Deutsche Forschungsgemeinschaft, Grant/Award Number: EXC 2075 - 39074001; National Science Foundation, Grant/Award Number: 1650042

## Abstract

Categorical moderators are often included in mixed-effects meta-analysis to explain heterogeneity in effect sizes. An assumption in tests of categorical moderator effects is that of a constant between-study variance across all levels of the moderator. Although it rarely receives serious thought, there can be statistical ramifications to upholding this assumption. We propose that researchers should instead default to assuming *unequal* between-study variances when analysing categorical moderators. To achieve this, we suggest using a mixed-effects location-scale model (MELSM) to allow group-specific estimates for the between-study variance. In two extensive simulation studies, we show that in terms of Type I error and statistical power, little is lost by using the MELSM for moderator tests, but there can be serious costs when an equal variance mixed-effects model (MEM) is used. Most notably, in scenarios with balanced sample sizes or equal between-study variance, the Type I error and power rates are nearly identical between the MEM and the MELSM. On the other hand, with imbalanced sample sizes and unequal variances, the Type I error rate under the MEM can be grossly inflated or overly conservative, whereas the MELSM does comparatively well in controlling the Type I error across the majority of cases. A notable exception where the MELSM did not clearly outperform the MEM was in the case of few studies (e.g., 5). With respect to power, the MELSM had similar or higher power than the MEM in conditions where the latter produced non-inflated Type 1 error rates. Together, our results support the idea that assuming unequal between-study variances is preferred as a default strategy when testing categorical moderators.

**KEYWORDS**
heterogeneity, location-scale modelling, meta-analysis, mixed effects, moderator, random effects

# 1 | INTRODUCTION

Meta-analysis is an indispensable technique for synthesizing results from various comparable studies. Key goals include determining the average effect size (Hedges & Pigott, 2001), quantifying the extent of heterogeneity in effects (Higgins & Thompson, 2002; Ioannidis et al., 2007), and evaluating whether study characteristics moderate the effect size (Hedges & Pigott, 2005). The study of statistical methods for meta-analysis is important because systematic reviews using meta-analytic techniques tend to be influential (DeGeest & Schmidt, 2010). For example, in psychological science, meta-analysis has been used to critically evaluate prominent theories (Hagger et al., 2016), provide nuanced perspectives on phenomena of interest (van Agteren et al., 2021), and suggest potential reasons for why replication studies fail (Klein et al., 2018). It is thus of paramount importance to use methods that provide accurate estimates and high-quality inferences.

Moderator variables are often included in meta-analytic models to investigate whether particular study characteristics (e.g., measurement instrument) explain differences in effect sizes and are typically viewed as a means to explain between-study heterogeneity (Thompson & Sharp, 1999). In a standard random-effects framework, the leftover, unexplained between-study variance is then captured by a heterogeneity parameter, $\tau^2$. Note that the term mixed-effects model (MEM) is often used in reference to a random-effects model with moderators, and we use both terms interchangeably throughout this paper. An implicit assumption in moderator analyses that has yet to be thoroughly examined is that $\tau^2$ is fixed across all subgroups for a categorical moderator, or across all values for a continuous moderator. That is, researchers commonly assume that, say, both subgroups of a dichotomous moderator will have the same value for $\tau^2$. This is analogous to the assumption of homoscedasticity in analysis of variance (ANOVA) designs. When this assumption does not hold true in the population and the subgroups of the moderator have unbalanced sample sizes, this can result in increased Type I error rates, conservative Type I error rates, and a loss in statistical power for moderator effects (Rubio-Aparicio et al., 2017).

One way of dealing with heterogeneous between-study variances with categorical moderators is by subgroup analysis (Borenstein & Higgins, 2013; Schoemann, 2016), wherein a separate random-effects model is fit to each level in the moderator and separate estimates of $\tau^2$ for each subgroup are obtained. Owing to their flexibility to examine multiple predictors within a single modelling framework, a popular alternative to subgroup analyses are mixed-effects meta-regression models (Thompson & Sharp, 1999). Standard mixed-effects meta-regressions, or MEMs, do not permit heterogeneous variances between subgroups, but Viechtbauer and López-López (2022) recently described a mixed-effects location-scale model (MELSM) in order to overcome this limitation, while Williams et al. (2021) introduced this model under a Bayesian framework. Using a MELSM, separate estimates of $\tau^2$ can be obtained for each level of the moderator (scale), which in turn yield modified estimates of so-called moderator effects (location). As we will show, allowing for separate heterogeneity estimates curbs the loss of power and distorted error rates when there are unequal between-study variances and unequal sample sizes in subgroups of a moderator. Unlike previous work on MELSMs, the intent of this work is not to introduce and describe the MELSM and its applications, but rather to understand whether it is preferable as a default strategy over a standard mixed-effects meta-regression.

## 1.1 | Analogy to the two independent samples *t*-test

To understand why the MELSM may be suitable as a default meta-analytic model, it is informative to consider the *t*-test. There is a rich and storied literature examining the *t*-test when the population variances of the two groups are heterogeneous (Bartlett, 1936; Murphy, 1967; Welch, 1938, 1947). When the sample sizes of the groups are balanced, inferences stemming from Student's *t*-test are generally unaffected by unequal group variances, save for extreme cases (Scheffé, 1959, Ch. 10). Conversely, when the variances of the groups and their respective sample sizes are both heterogeneous, the resulting statistics can be drastically biased and lead to invalid inferences—the well-known Behrens-Fisher problem. To overcome these issues, Welch's *t*-test is commonly used and has even been recommended to be the default because it adjusts the degrees of freedom according to group sample sizes and variances (Delacre et al., 2017).

In the same vein, Rubio-Aparicio et al. (2017); Rubio-Aparicio et al. (2020) found that in subgroup analysis, using a pooled estimate of $\tau^2$ (i.e., assuming equal between-study variances) yielded inferences similar to those yielded when using separate estimates (i.e., assuming unequal between-study variances) if the subgroup sample sizes were at least roughly equal. They concluded that pooling estimates for $\tau^2$ was to be preferred in most scenarios—a sentiment that is echoed in the broader meta-analysis literature (e.g., Borenstein et al., 2009; Viechtbauer, 2010). We view this hesitation to assume unequal between-study variances as unwarranted because, as we will show, little is lost by simply assuming an unequal variance model, but there are potentially costly consequences to assuming equal variances.

### 1.2 | Overview

The outline of this paper is as follows. We first describe the standard MEM and the MELSM. We then describe two popular tests of moderator effects and highlight the importance of $\tau^2$ in estimating the moderator coefficients and their respective standard errors. This section also clarifies how moderator tests are essentially weighted tests of mean differences. Next we present extensive simulation studies comparing the error rates and statistical power of moderator tests using the MELSM to those of obtained using a classical MEM. We conclude with an applied illustration of the MELSM and a brief discussion.

## 2 | MIXED-EFFECTS MODELS

In meta-analysis, researchers are routinely faced with the choice of using either a fixed-effects or random-effects model. The former provides inferences on the effect sizes of the observed studies or a set of identical studies (i.e., conditional inference). The latter yields inferences on the parameters of a population of studies, of which the observed set is assumed to be a random sample (i.e., unconditional inference; Hedges & Vevea, 1998). In addition to sampling variability, the random-effects model considers the possibility of heterogeneous true effect sizes that arise as a result of differences in study characteristics (e.g., design, population, environmental factors; DerSimonian & Laird, 1986; Hedges, 1992). The main consideration in deciding between these two models is the inferential goal. In practice, analysts often wish to generalize beyond the observed studies to the parent population of studies, and thus random-effects models are more often employed.

### 2.1 | Standard MEM

When a moderator is included in a random-effects model to explain part of the variance in the observed effect sizes, it becomes a MEM. Additionally, for categorical moderators, the inferential goal is usually to contrast between two or more group average effects. In this case, the MEM can be understood as follows. Suppose we have observed $j = 1, \ldots, k_i$ independent effect sizes in each of $i = 1, \ldots, p$ groups, $y_{ij}$, and that the total number of observed effect sizes is $k = \sum_i k_i$. For simplicity, suppose that $p = 2$ (i.e., two groups). Then the standard MEM is given by

$$y_{ij} = \theta_{ij} + \varepsilon_{ij}, \tag{1}$$

$$\theta_{ij} = \mu_i + u_{ij}, \tag{2}$$

$$\mu_i = \beta_0 + \beta_1 x_{ij}, \tag{3}$$

$$u_{ij} \sim \mathcal{N}(0, \tau^2), \tag{4}$$

$$\varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma_{ij}^2\right), \tag{5}$$

where $\theta_{ij}$ is the true effect size of the *j*th study in the *i*th group, $\beta_0$ is the average effect for the first group, $\beta_1$ captures the difference in average effect between the first and second groups, and $x_{ij}$ is a dummy variable indexing the group to which each observed effect size belongs.

The terms $\varepsilon_{ij}$ and $u_{ij}$ are the within- and between-study errors, respectively. The within-study errors are assumed to be normally distributed with zero mean and variance $\sigma_{ij}^2$. Each $\sigma_{ij}^2$ is assumed to be known in advance and is calculated according to the type of effect size index being investigated (Hedges & Olkin, 1985). The between-study errors are similarly assumed to be normally distributed with zero mean and variance $\tau^2$. This parameter captures the variation in the true effect sizes, $\theta_{ij}$, that is not explained by the moderators and can be estimated in a number of ways (see Langan et al., 2017, appendix C). In a frequentist framework, perhaps the most common estimator of $\tau^2$ is the restricted maximum likelihood (REML; Patterson & Thompson, 1971) estimator, although $\tau^2$ is also often estimated using Bayesian techniques (e.g., Higgins et al., 2009). Importantly, in the MEM, $\tau^2$ does not carry a subscript, and hence the assumption of equal variances among groups of a categorical moderator is always made.

This assumption cannot be overlooked because the role of $\tau^2$ is crucial in calculating and testing moderator effects. The way in which the between-study variance affects the outcome of moderator tests is evident in the estimating equations for $\beta_0$ and $\beta_1$ and their standard errors. These equations are solved using weighted least squares (Viechtbauer, 2010), where the weights are calculated based, in part, on the between-study variances. When there are two groups, the equations for testing a moderator effect (i.e., $\beta_1$) bear a striking resemblance to those of a weighted Student's *t*-test. For instance, the well-known *Q* statistic with two groups can be expressed as (Hedges & Pigott, 2005)

$$Q = \frac{\left(\bar{y}_1 - \bar{y}_2\right)^2}{w_1^{-1} + w_2^{-1}}, \tag{6}$$

where, just as for the *t* statistic in an independent two-sample *t*-test, the numerator reflects the observed mean difference in the outcome between groups and the denominator captures the pooled variance of the numerator (see Tests for Categorical Moderators for details). In the case of more than two groups, testing a moderator effect corresponds to a weighted ANOVA (Borenstein et al., 2009). A key assumption of standard ANOVA methods is that the within-group variances are equal, and this assumption is also commonly made in meta-analysis, but with respect to the between-study variance.[1] However, recall that our aim is to understand whether this assumption is warranted.

## 2.2 | Mixed-effects location-scale model

To accomplish the goal of estimating and incorporating heterogeneous variances into tests of moderator effects, we propose using a MELSM (Viechtbauer & López-López, 2022; Williams et al., 2021). This technique was first introduced outside of meta-analysis[2] (Hedeker et al., 2008, 2012), and has also been studied under the terms *doubly hierarchical model* (Lee & Nelder, 2006) and *distributional regression* (Bürkner, 2018).

[1] The within-study variances in a meta-analysis are heterogeneous by default and are typically considered to be known in advance.

[2] Note the "location-scale" term arose because of the inclusion of a random effect for both the location and level-1 variance models in a non-meta-analytic MEM, but we use it here to refer to a mixed-effects meta-regression that allows moderators for the location and level-2 variance.

This model is formulated similarly to the standard MEM in (1), but an additional component is stipulated to permit moderators to influence the between-study variance. When a categorical moderator is included, it results in group-specific estimates for the between-study variance, $\tau_i^2$. If we continue with our simplifying assumption that $p = 2$, then $\tau_i^2$ can be modelled through the log-linear equation

$$\tau_i^2 = \exp(\gamma_0 + \gamma_1 x_{ij}), \tag{7}$$

where $\gamma_0$ captures the between-study heterogeneity for the reference group $\tau_1^2$, $\gamma_1$ captures the difference between $\tau_1^2$ and $\tau_2^2$, and $x_{ij}$ is the same group-indicator variable used in (3). The right hand side of (7) is exponentiated to ensure only positive values are estimated for $\tau_i^2$. If only an intercept is included in predicting between-study variance, then this results in an estimate of the average heterogeneity across all $k$ studies, or the usual estimate of $\tau^2$. In this sense, the MEM can be considered a special case of the MELSM. Additionally, the coefficients $\gamma_0$ and $\gamma_1$ are typically estimated using either REML (Viechtbauer, 2021) or Bayesian methods (Williams et al., 2021).

Critically, because the MELSM allows each group to have its own value for $\tau_i^2$, it follows that a different set of weights is used in estimating the coefficients $\beta_0$ and $\beta_1$ (see Hedges & Pigott, 2005, p. 442), as well as their respective standard errors. Thus, just as the moderator tests using MEMs are akin to weighted tests of group means with equal variances, moderator tests using the MELSM can be likened to weighted tests of group means with unequal variances.

## 2.3 | Tests for categorical moderators

There are several popular methods in meta-regression for testing moderator effects. Of these, perhaps the most widely used is the Wald-type $z$-test (Borenstein et al., 2009). This test, however, is known to not adequately control the Type I error rate because it ignores the uncertainty involved in estimating the between-study variance (Knapp & Hartung, 2003; Viechtbauer et al., 2015). As we will see, this in turn affects the estimates and standard errors of the moderator effect. An alternative method originally proposed by Knapp and Hartung (2003) amends this deficiency by applying a correction factor to the variance of the moderator effect (e.g., Rubio-Aparicio et al., 2020). For these reasons, we examine how the weights resulting from the MEM and the MELSM affect both the $z$-test and Knapp and Hartung (KH) method.

To crystallize our ongoing analogy between classical tests for mean differences and tests of moderator effects, the expressions we use are for the $t$-test and ANOVA-like counterparts of the $z$-test and KH method, but they are equivalent to meta-regression equations when there are two groups under consideration. That is, the omnibus $Q$-test is equivalent to the $z$-test for meta-regression coefficients when there are two levels in a categorical moderator (cf. Hedges & Pigott, 2005, pp. 434–435 and pp. 442–443). Similarly, the KH method for regression coefficients (Knapp & Hartung, 2003) is equivalent to the $F$-test described in Hartung et al. (2001) in the case of two groups. Describing the tests in these forms will help us clearly show that, in fact, assuming unequal between-study variances in meta-analyses is no different than assuming unequal variances in classical tests of mean differences. Importantly, the latter is already routine practice in psychological science.

### 2.3.1 | Q statistic

In this section and the one following it, we focus on frequentist estimation of moderator effects, but it should be mentioned that the same logic we use would similarly hold for Bayesian estimation (Williams et al., 2021). Testing a moderator effect entails computing a series of *weighted* means. The first step is to calculate the weighted mean effect size for each group, where the weights are given by the inverse of the total variance in $y_{ij}$. Under the standard MEM, we have the weights

$$w_{ij}^{*} = \frac{1}{\sigma_{ij}^{2} + \tau^{2}}. \tag{8}$$

Under the MELSM, the weights are given by Williams et al. (2021)

$$w_{ij}^{**} = \frac{1}{\sigma_{ij}^{2} + \tau_{i}^{2}}. \tag{9}$$

As mentioned earlier, the key difference in the weights used by the MEM and the MELSM is that the former assumes a common $\tau^2$ across all $i$ groups, whereas the MELSM can incorporate group-specific heterogeneity parameters, $\tau_i^2$. Using these weights, the mean for each group can be calculated as

$$\overline{y}_i = \frac{\sum_{j=1}^{k_i} w_{ij} y_{ij}}{\sum_{j=1}^{k_i} w_{ij}}. \tag{10}$$

Note that here and in what follows, we use the term $w_{ij}$ in a generic sense to indicate either $w_{ij}^{*}$ or $w_{ij}^{**}$, depending on whether equal or unequal variances are assumed. If there are no differences in between-study variances ($\tau_1^2 = \cdots = \tau_p^2 = \tau^2$), then the weights in (9) are equal to those in (8). Additionally, the $\tau_i^2$ are not usually known in practice and so are usually replaced with their estimates, $\hat{\tau}_i^2$.

The $i$th group mean is assumed normally distributed with mean $\mu_i$ and variance $\sigma_i^2 = \left(\sum_{j=1}^{k_i} w_{ij}\right)^{-1}$. These variances can in turn be used to obtain the weighted grand mean of effect sizes across groups

$$\overline{y} = \frac{\sum_{i=1}^{p} w_i \overline{y}_i}{\sum_{i=1}^{p} w_i}, \tag{11}$$

$$w_i = 1/\sigma_i^2. \tag{12}$$

With $\overline{y}_i$ and $\overline{y}$ in hand, it is possible to test the hypothesis $H_0$: $\mu_1 - \mu_2 = 0$ by taking the test statistic

$$Q = \frac{\left(\overline{y}_1 - \overline{y}_2\right)^2}{w_1^{-1} + w_2^{-1}}, \tag{13}$$

or, in the case of testing $p > 2$ group means,

$$Q = \sum_{i=1}^{p} w_i \left(\overline{y}_i - \overline{y}\right)^2, \tag{14}$$

and referring it to a $\chi^2$ distribution with $(p-1)$ degrees of freedom. If the significance level of the test is $a$ and $c_a$ denotes the $(1-a)$ quantile of the central $\chi^2$ distribution, then $H_0$ can be rejected if $Q > \mathrm{c}_{\mathrm{a}}$.

### 2.3.2 | Knapp and Hartung method

The KH method builds on the omnibus $Q$-test by applying an adjustment that accounts for the uncertainty in estimating the between-study variance. This adjustment has been found to ameliorate the Type I error rate of the $Q$-test in MEMs (Rubio-Aparicio et al., 2020) and results in the $F$-statistic given by

$$F = \frac{Q/(p-1)}{Q_{\mathrm{W}}/(k-p)}, \tag{15}$$

where $Q_\mathrm{w}$ is defined as

$$Q_\mathrm{w} = \sum_{i=1}^{p} \sum_{j=1}^{k_i} w_{ij} \left( y_{ij} - \overline{y}_i \right)^2 . \tag{16}$$

The $F$-statistic can be used to test the hypothesis $H_0 : \mu_1 = \cdots = \mu_\mathrm{p}$ by referring it to an $F$-distribution with $(p-1)$ and $(k-p)$ degrees of freedom. If the significance level of the test is $a$ and $c_a$ denotes the $(1-a)$ quantile of the central $F$-distribution, then $H_0$ can be rejected if $F > c_a$.

There are two important details to observe here. First, the expressions used for the $Q$- and $F$-tests are in effect the same as those used in a $t$-test and ANOVA, but they use *weighted* forms of the group and grand means; thus, assuming unequal $\tau_i^2$ is equivalent to assuming unequal variances in a $t$-test or ANOVA. Second, and relatedly, the weights $w_{ij}^*$ and $w_{ij}^{**}$ are fundamentally tied to the between-study variances. It follows that if groups in a moderator truly have different $\tau_i^2$, then the weights using a pooled value will be incorrect, as will the test statistic. This holds implications for both the Type I error and power of moderator tests, both of which are investigated in the section Simulation Studies.

### 2.3.3 | Motivating example

To highlight the impact of the between-study variance on the resulting test statistic, we calculated the $Q$- and $F$-values for a hypothetical set of $k = 40$ studies where each study belonged to one of $p = 2$ groups. Following Williams et al. (2021), we assumed all $\sigma_{ij}^2$ and $\tau_i^2$ were known, with the $\sigma_{ij}^2$ fixed at .04 (roughly the variance of a standardized mean difference (SMD) with $n = 50$). The heterogeneity parameter for the first group $\tau_1^2$ was held constant at .05 but varied for the second group from .05 to .25 in steps of .01. We further varied the proportion of total studies belonging to each group such that the proportion of the 40 studies in the first group ranged from .5 to .9 in .1 increments. Lastly, we calculated the $Q$- and $F$-statistics,[3] using weights that used either the heterogeneous $\tau_i^2$-values or a pooled $\tau^2$-value. The denominator for the $F$-statistic was set to $Q_w = .67(k-p)$, which corresponds to a “moderate” degree of within-group heterogeneity (Hedges & Pigott, 2005). Pooled values of $\tau^2$ were obtained by taking the arithmetic mean of the $\tau_i^2$ across all studies. The main idea here is that if the between-study variances influence the value of the test statistics, then the values for the $Q$- and $F$-statistics should differ depending on whether equal or unequal $\tau_i^2$ are assumed.

Figure 1 includes the results for this example. The $y$-axes denote the ratio of the test statistic using the pooled value of $\tau^2$ ($Q_\mathrm{Equal}$ and $F_\mathrm{Equal}$) to the test statistic using the heterogeneous values for $\tau_i^2$ ($Q_\mathrm{Unequal}$ and $F_\mathrm{Unequal}$). Thus, a ratio of 1 indicates that the test statistics are equal, and positive values indicate that the statistic using equal $\tau_i^2$ is greater than the one using unequal $\tau_i^2$. The $x$-axes display the between-study variance for the second group, $\tau_2^2$, and colours differentiate the number of studies in the second group. In line with the classical literature on the $t$-test and ANOVA, the resulting test statistics are invariant to heterogeneous between-study variances under balanced sample sizes. They are also invariant under imbalanced sample sizes when the between-study variances are equal and known, as in this example. The test statistics differed, however, when both the between-study variances and sample sizes were heterogeneous across the two groups. In our setup, greater disparities in either factor led to the statistics that used equal between-study variances being larger than those that used heterogeneous between-study variances. In the worst case with $k_2 = 4$ and $\tau_2^2 = .25$, the $Q_\mathrm{Equal}$ and $F_\mathrm{Equal}$ statistics were 1.42 times larger than their unequal counterparts.

Despite the simplicity of this example, it is clear that when heterogeneous variances and sample sizes go unaccounted for when testing moderators, the resulting test statistic can lead to overconfidence in the strength of the effect. Parallel to how Welch's proposed versions of the $t$-test and ANOVA deal with such scenarios, we too propose that unequal (between-study) variances should be assumed by default when

[3] To provide analytical results in this example we used the equations given in the section Statistical Power.

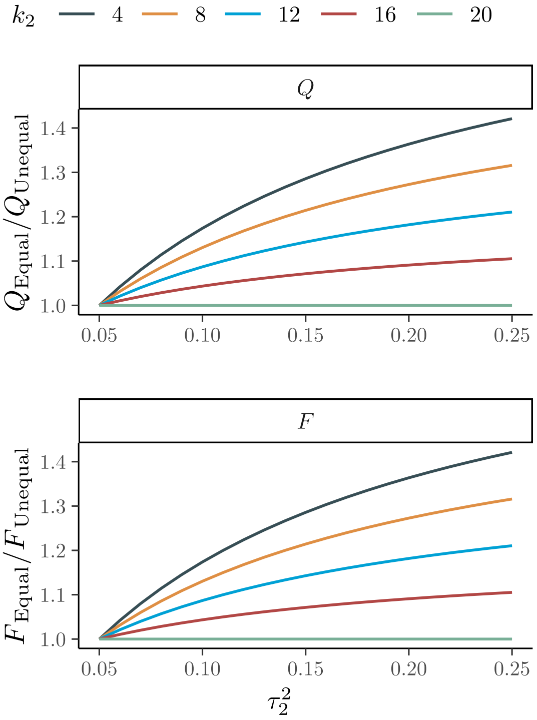

**FIGURE 1** Motivating example of how the $Q$- and $F$-statistics vary according to whether equal or unequal between-study variances are assumed (see main text for details). As the imbalances grow between groups in between-study variances and sample sizes, so too does the ratio of the test statistics.

testing categorical moderators. The motivation behind this is that assuming unequal between-study variances will result in the appropriate test statistics regardless of whether or not they are indeed equal, but assuming them to be equal will be appropriate only if the population between-study variances are truly equal. In the next section, we study the behaviour of the MELSM when testing moderators under various realistic scenarios to ensure its suitability as a default method.

## 3 | SIMULATION STUDIES

Because we are proposing that unequal between-study variances should be assumed by default when testing categorical moderators, it is of interest to examine how the MELSM performs relative to the standard MEM in situations where the variances are equal and when they are not. Accordingly, we conducted two Monte Carlo simulation studies comparing the performance of the two models in terms of Type I error and statistical power. Specifically, we compared these two quantities with respect to the $Q$-statistic defined in (13) and the $F$-statistic defined in (15). Unlike the preceding example, the use of Monte Carlo simulations allows us to examine the empirical performance of the $Q$- and $F$-tests in a more realistic setting because in practice the between- and within-study variances are not fixed and known in advance. We chose SMDs as the effect size measure and a two-level categorical variable as a moderator.

### 3.1 | Method

Following standard procedure in psychology, we assumed that each simulated study compared an experimental group to a control group on an arbitrary quantitative outcome (e.g., López-López et al., 2014;

Sánchez-Meca & Marín-Martínez, 2008; Viechtbauer et al., 2015). Under a MEM, the population SMD for each study was defined as

$$\theta_{ij} = \frac{E_{ij} - C_{ij}}{S_{ij}}, \tag{17}$$

where $E_{ij}$ and $C_{ij}$ were the population means for the experimental and control groups, respectively, and $S_{ij} = 1$ was the pooled within-study standard deviation. The population distribution for the $\theta_{ij}$ was taken to be $\mathcal{N}(\mu_i, \tau_i^2)$, where the subscript denotes the $i$th group. The specific values for $\mu_i$ and $\tau_i^2$ were systematically varied in each simulation study and are given in what follows. For the experimental and control groups, we defined the population distributions as $\mathcal{N}(\theta_{ij}, 1)$ and $\mathcal{N}(0, 1)$, respectively. Using these definitions, a study was simulated by first drawing a $\theta_{ij}$ and then drawing $n_{E_{ij}} = n_{C_{ij}}$ variates from the latter two distributions and estimating the SMD in the usual way (Hedges & Olkin, 1985).

To be clear, the sample sizes $n_{E_{ij}}$ and $n_{C_{ij}}$ respectively refer to the sample size of the experimental and control groups *within* studies and are not to be confused with the number of studies in each subgroup of a categorical moderator variable, $k_i$. The within-study sample sizes were set to $n_{E_{ij}} = n_{C_{ij}} = g$, where $g$ is a gamma variate divided by two and rounded to the nearest integer. Gamma variates were drawn from a gamma distribution with shape equal to .65 and scale equal to 165, truncated to values greater than or equal to thirty. Similar procedures were used to generate within-study sample sizes consistent with those commonly observed in social-behavioural studies (Brannick et al., 2019; Dahlke & Wiernik, 2020).

Using the described procedure, we generated a collection of $k = k_1 + k_2$ studies for each simulation trial and conducted moderator tests using both the MEM and the MELSM. All models were fit using REML estimation. Both the study sample size ($k_1 = 20$) and between-study variance ($\tau_1^2 = .2$) were held constant for the first group but varied for the second group. The value of $\tau_2^2$ varied from .05 to .4 in steps of .05, and values of 5, 10, 20, 30, and 40 were examined for $k_2$. These values were chosen because they represent empirical values of $\tau^2$ observed in psychological meta-analyses (van Erp et al., 2017). Additionally, the classical literature examining tests with unequal group variances focuses not on the raw values of variances and sample sizes, but on their ratios (Bartlett, 1936; Murphy, 1967; Scheffé, 1959; Welch, 1938), and in line with this convention, the study sample sizes and variances we considered approximately corresponded to the following ratios: (a) $k_1/k_2 = \left(\frac{1}{4}, \frac{1}{2}, 1, 1.5, 2\right)$ and (b) $\tau_1^2/\tau_2^2 = \left(\frac{1}{4}, \ldots, 2\right)$. Using ratios rather than raw values further lends intuitive credibility regarding how the group variances and number of studies affect Type I error and power in situations outside of those considered here.

Our expectations for these simulation studies were as follows. As with traditional tests of mean differences, we expected that for each moderator test, the Type I error rate and power rate would be approximately equivalent between the MEM and the MELSM when the sample sizes were balanced and/or the between-study variances were equal. When the group sizes were imbalanced, we expected two general trends to emerge as a result of assuming equal between-study variances: when the group with a larger sample size simultaneously had a smaller between-study variance, then assuming equal variances would lead to a higher Type I error rate, and conversely, when the larger group had a larger between-study variance, then assuming equal variances would result in a conservative Type I error rate. These scenarios were also expected to yield invalid power rates because the rejection rate would be too high or too low. We expected, however, that assuming unequal between-study variances in these scenarios would result in a nominal error rate and valid statistical power (i.e., no increased/decreased power at the cost of an inflated/conservative rejection rate). In sum, we anticipated potentially serious costs in terms of Type I error and power when using the MEM, but not the MELSM.

For each simulation study, the total number of conditions was 2 ($Q$- or $F$-test) × 2 (MEM or MELSM) × 5 ($k_2$) × 8 ($\tau_2^2$) = 160, each of which was simulated 10,000 times. All simulations were conducted in the R environment (R Core Team, 2021) using the **metafor** package to fit the meta-regression models (Viechtbauer, 2010). All code used in this article is openly available on the Open Science Framework at https://osf.io/gs6uw/.

## 3.2 | Type I error

To examine how well the MELSM controls Type I error of the $Q$- and $F$-tests relative to the MEM, we set $\beta_0 = \beta_1 = 0$, or, in terms of group means, we set $\mu_1 - \mu_2 = 0$. For both moderator tests, the $a$ level was set to .05. Further, we considered "stringent" and "liberal" criteria for assessing adequate control of Type I error that respectively corresponded to the intervals [.046, .054] and [.043, .057]. These intervals were determined by computing the critical values of a two-tailed test that a population proportion was equal to .05 under significance levels of .05 and .001 and with a sample size of 10,000. In other words, for each condition, values outside of the intervals would result in a rejection of the null hypothesis that the Type I error rate was equal to .05 at the given significance level.

The results are presented in Figure 2. Type I error rates for both the $Q$- and $F$-tests are plotted as a function of the between-study heterogeneity for the second group ($\tau_2^2$), whether the variances were treated as equal (MEM) or unequal (MELSM), and the number of studies in the second group ($k_2$). The dashed white line denotes the nominal error rate $a = .05$, the light grey area region represents the region between [.043, .057], and the dark grey shaded area represents the region in the range [.046, .054]. The colour of the solid lines distinguishes between the number of studies included in the second group.

The rejection rates were in line with our expectations and largely mirror those in the classical $t$-test and ANOVA literature. When the between-study variances were equal ($\tau_1^2 = \tau_2^2 = .2$), the Type I error was roughly the same between the MEM and MELSM, regardless of study sample size. This was also the case across all values for $\tau_2^2$ when the study sample sizes were equal ($k_1 = k_2 = 20$). The only time the MEM clearly controlled the Type I error rate better than the MELSM was when $k_2 = 5$ *and* the between-study

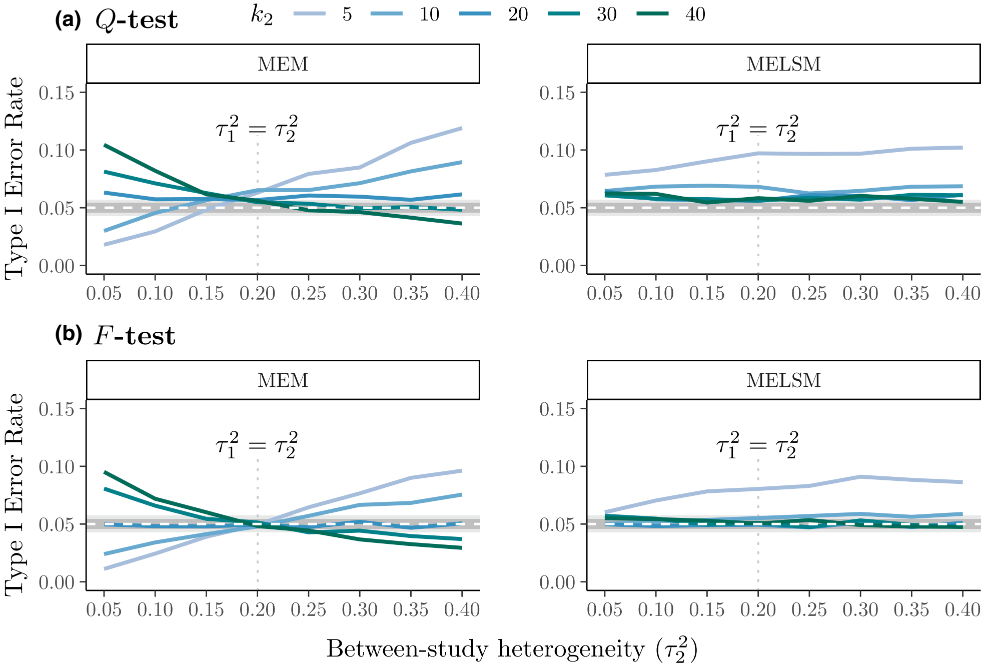

FIGURE 2 Type I error rates for $Q$-test (panel a) and $F$-test (panel b) under MEM (equal $\tau^2$ assumed) and MELSM (unequal $\tau_i^2$ assumed). The white dashed line denotes .05, the light grey area spans [.043, .057], and the dark grey area spans [.046, .054]. For both moderator tests, the MEM only resulted in a well-controlled error rate when either the $\tau_i^2$ were truly equal or when the sample sizes were balanced. when there were unequal $\tau_i^2$ and unbalanced sample sizes, but the MEM was used, the error rate was inflated or excessively conservative. Using the MELSM, however, mostly resulted in a well-controlled type I error rate regardless of whether the variances and sample sizes were equal (except when $k_2 = 5$). Although both moderator tests had better overall type I error under the MELSM, only the $F$-test had error rates consistently inside of the [.046, .054] bounds.

variances were equal. In other words, in all but one case, the MELSM was on par with the MEM in adequately controlling the Type I error even when the assumption of equal between-study variances was true in the population. The similar rejection rates between the MEM and the MELSM under balanced sample sizes and equal $\tau_i^2$ lends strong support to our view that almost nothing is lost by assuming unequal between-study variances by default. On the other hand, there may be serious ramifications of assuming equal between-study variances when they are unequal and are coupled with imbalanced sample sizes.

For both the *Q*- and *F*-test, when the between-study variances and sample sizes simultaneously differed between groups but equal $\tau_i^2$ were assumed, the Type I error rate deviated from the nominal rate. In particular, when the larger group (e.g., $k_1 > k_2$) also had a smaller between-study variance (e.g., $\tau_1^2 < \tau_2^2$), the error rate *increased* above the nominal level. In contrast, when the larger group had a larger between-study variance (e.g., $\tau_1^2 > \tau_2^2$), the error rate *decreased* below .05. In the most extreme cases, the MEM had a Type I error rate of .02 ($k_2 = 5, \tau_2^2 = .05$) and .12 ($k_2 = 5, \tau_2^2 = .4$), well outside the bounds of the wider [.043, .057] interval.

When unequal between-study variances with the MELSM were assumed, the Type I error rate was mostly well maintained across all conditions and regardless of moderator test, relative to the MEM, which assumed equal between-study variances. Compared against the liberal and stringent intervals, however, the *Q*-test was expectedly worse than the *F*-test in controlling the Type I error. For the *Q*-test, the error rates in the majority of the conditions fell outside of the wider [.043, .057] bounds. Even so, the error rates were still predominantly closer to nominal than those of the *Q*-test using the MEM. Meanwhile, the *F*-test achieved a much better Type I error rate, with most conditions achieving an error rate within the wide interval of [.043, .057] and many falling between the stricter interval of [.046, .054]. As mentioned earlier, the only cases where the moderator tests assuming unequal $\tau_i^2$ clearly performed worse than their equal between-study variance counterparts occurred when $k_2 = 5$, though it is worth pointing out that all methods did relatively poorly in such conditions. In all other conditions, the Type I error rates under the MELSM were as good as or better than those obtained using the MEM, with the *F*-test achieving the nominal error rate. These findings fortify the idea that the resulting inferences of moderator tests can be seriously compromised when equal between-study variances are assumed, but inferential integrity can be preserved (at least in part) by assuming unequal $\tau_i^2$.

## 3.3 | Statistical power

To assess the empirical statistical power rates of the MEM and the MELSM with respect to the two moderator tests, we followed nearly the exact same procedure as in the previous simulation study, but we set $\beta_1 = .5$, or $\mu_2 - \mu_1 = .5$.[4] Such a moderator effect can be regarded as a representing a "medium" effect size for a SMD (Cohen, 1992).

Figure 3 displays the results for $\tau_2^2 = (.05, .2, .4)$, corresponding to ratios of $\tau_1^2/\tau_2^2 = \left(\frac{1}{4}, 1, 2\right)$. Power rates are plotted as a function of studies in the second group ($k_2$), heterogeneity in the second group ($\tau_2^2$), whether the variances were treated as equal, and the moderator test. The dashed horizontal grey line denotes a power rate of .8. Additionally, we excluded conditions where the Type I error rate for the MEM was greater than the error rate for the MELSM because they could misleadingly be interpreted as achieving higher power. That is, we omitted conditions where the MEM produced higher power rates than the MELSM but only achieved higher power because they also had a higher Type I error rate. The complete set of results are included in the Appendix (Table A2). Similar to the Type I error, the power rates between the MEM and the MELSM were nearly identical with balanced sample sizes and equal between-study variances, regardless of the moderator test. The lack of differences between the MEM and

[4] We additionally conducted simulations for $\beta_1 = .2$ and $\beta_1 = .8$, but the results were qualitatively the same as those presented in what follows. Thus, they are omitted here but can be found in the Appendix (Tables A3 and A4).

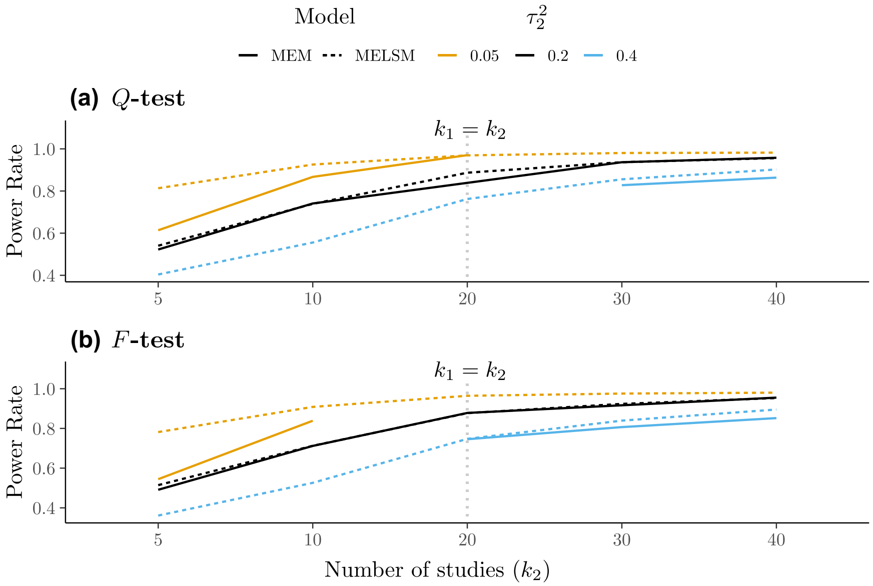

FIGURE 3 Empirical power rates for the $Q$-test (Panel a) and $F$-test (Panel b) for $\beta_1 = 0.5$ and $\tau_1^2 = 0.2$. Power rates for conditions that exceeded acceptable Type I error bounds are excluded. Solid lines denote the MEM with equal $\tau_i^2$ and dashed lines denote the MELSM with unequal $\tau_i^2$. The power rates for both models were similar under balanced sample sizes and equal between-study variances, but there were differences otherwise. Particularly, the moderators tests under the MEM produced higher or lower power relative to the MELSM, but in conditions where the tests also had Type I error rates that were above or below the nominal 0.05. In contrast, the moderator tests under the MELSM produced valid power rates across more conditions insofar as they did not incur any costs in terms of Type I error.

the MELSM in this scenario again supports our claim that there are essentially no drawbacks to assuming unequal $\tau_i^2$ by default.

With respect to situations where the sample sizes were unbalanced and the between-study variances were simultaneously unequal, power rates diverged between the MEM and the MELSM. Most notably, in conditions where the MELSM produced as good or better Type I error rates than the MEM, the MELSM also resulted in as good or better power rates, regardless of moderator test. For example, for the $F$-test, when $k_2 < 20$ and $\tau_2^2 = .4$, the MEM produced an inflated Type I error rate and thus did not produce valid power rates. In contrast, when $k_2 \geq 20$, the MEM had valid power rates, but they were either equal to or lower than the power rates for the MELSM. Similar conclusions can be drawn when considering different values for $k_2$ and $\tau_2^2$. Together, these findings add support to our argument that little is lost by assuming heterogeneous between-study variances when testing moderator effects with regard to statistical power, but there may be a lot to gain.

## 3.4 | Analytical solution to power

Beyond simulation, the power rates of these tests can be assessed analytically, albeit at the cost of making a few more assumptions. Because simulation-based power analyses can be computationally demanding and time consuming, an analytic solution can be particularly useful if researchers wish to conduct a power analysis prior to carrying out a moderator test. To calculate statistical power for the omnibus $Q$-test, values must be assumed for $k$ and $p$ and for each $\mu_i$, $\tau_i^2$, $\sigma_{ij}^2$, and $k_i$. Once these values have been chosen, power can be calculated as (Hedges & Pigott, 2005)

$$1 - \chi^2(c_\alpha | p - 1; \lambda), \tag{18}$$

where $\chi^2(|)$ denotes the cumulative distribution function (CDF) of a $\chi^2$ distribution with $(p-1)$ degrees of freedom and noncentrality parameter $\lambda$. Recall that in our formulation, $p$ refers to the number of groups in the categorical moderator. The term $c_a$ is the critical value for the central $\chi^2$ distribution used for the $Q$-test in (13). The noncentrality parameter $\lambda$ is given by

$$\lambda = \sum_{i=1}^{p} w_i(\mu_i - \mu)^2 \tag{19}$$

$$\mu = \frac{\sum_{i=1}^{p} w_i\mu_i}{\sum_{i=1}^{p} w_i}. \tag{20}$$

The expression to calculate power for the KH method is similar to that of the $Q$-test and can be computed as (Hartung et al., 2001)

$$1 - F(c_\alpha \mid p-1, k-p; \lambda), \tag{21}$$

where $F(|)$ denotes the CDF of an $F$-distribution with $(p-1)$ and $(k-p)$ degrees of freedom and noncentrality parameter $\lambda$. The term $c_a$ is the critical value for the central $F$ distribution used for the $F$-test in (15). The noncentrality parameter for the $F$-test is also similar to that of the $Q$-test and is defined as

$$\lambda = \frac{\sum_{i=1}^{p} w_i(\mu_i - \mu)^2}{\sum_{i=1}^{p}\sum_{j=1}^{k_i} w_{ij}\left(\theta_{ij} - \mu_i\right)^2}, \tag{22}$$

where $\theta_{ij}$ is the population effect size for the $i$th study in the $j$th group. Here, along with the values that must be assumed for the $Q$-test, values must also be inserted for $\theta_{ij}$. Because choosing values for the $\theta_{ij}$ can be particularly difficult, it has been suggested that values of $.33(k-p)$, $.67(k-p)$, and $(k-p)$ can be used for the denominator in (22), which correspond respectively to small, moderate, and large degrees of within-group heterogeneity (Hedges & Pigott, 2001).

To aid researchers in sample size planning, we provide the analytical solutions to power for a moderator effect of .5 under the same conditions as the power simulation but varied $k_2$ in steps of 1 from 5 to 40. For both tests, we set the within-study variances $\sigma_{ij}^2$ to .2, which is approximately the sampling variance of a SMD with $n = 50$. For the $F$-test, we computed power for various degrees of within-group heterogeneity. In particular, $F_1$, $F_2$, and $F_3$ coincide with the $F$-test under a small, moderate, and large degree of within-group heterogeneity, respectively.

The power rates for the MELSM are plotted in Figure 4. Here, power is plotted as a function of studies in the second group ($k_2$), the between-study heterogeneity in the second group ($\tau_2^2$), and moderator test. The pattern of power rates between the MEM and the MELSM was similar to the one discussed earlier, and so power rates for the former are omitted. Regarding the between-study variance, $\tau_2^2$, power is higher across all conditions with smaller values. A similar statement can be made for the degree of within-group heterogeneity with respect to the $F$-test. Perhaps intuitively, the $F$-tests with small and moderate within-group heterogeneity (i.e., $F_1$ and $F_2$) were more powerful than the $Q$- and $F_3$-tests across all conditions.

### 3.4.1 | Mean squared error and standard error

We conducted two supplementary comparisons between the MEM and the MELSM. In the foregoing empirical power simulation, we further collected (1) the mean squared error (MSE) for the moderator effect ($\beta_1$) in each condition and (2) the average standard error of the moderator effect in each condition.

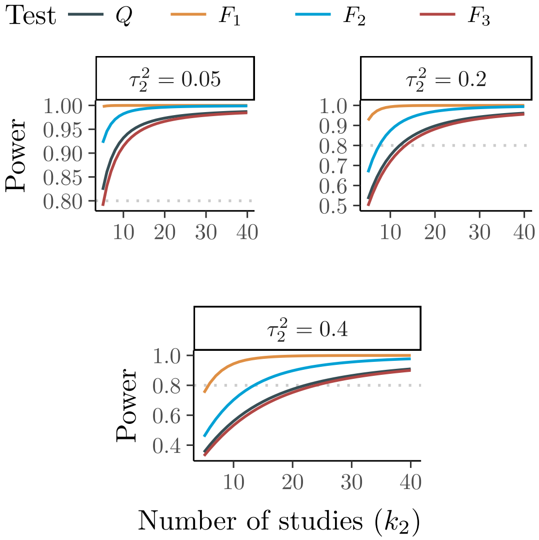

FIGURE 4 Analytical power for a moderator effect of .5 for the $Q$-test and $F$-test under fixed assumptions (see main text for details) for the MELSM. The lines corresponding to $F_1$, $F_2$, and $F_3$ denote the $F$-test with small, moderate, and large amounts of within-group heterogeneity, respectively. The $Q$-test and $F_3$-test most closely matched the observed power rates.

This was done to investigate discrepancies in the quality of the estimates for the effect itself and the respective standard error between the MEM and the MELSM.

The results can be viewed in Figure 5. Panel a presents the MSE in each condition, and Panel b displays the corresponding (average) standard errors. The grey lines denote $\tau_2^2$ values spanning from .05 to .4 in increments of .05 (dark to light). Solid and dashed lines differentiate between the MEM and the MELSM, respectively. Although the standard errors differ depending on whether the $Q$-test or $F$-test[5] is used for the moderator effect, their results were so similar that they are only shown for the former. The MSE was practically identical regardless of which model was used to estimate the moderator effect. Point estimates of the moderator effect were evidently unaffected by whether equal or unequal $\tau_i^2$ were assumed, whereas this choice affected the average standard error. When the number of studies, and thus the between-study variances, was unequal, the standard errors were either too large or too small, depending on the condition. Although the between-study variances indeed play a role in estimating the moderator effect, it seems that it is their effect on the standard errors that is responsible for the discrepancies between the MEM and the MELSM observed in the aforementioned simulation studies.

## 3.5 | Extension to three groups

Naturally, one may wonder whether the previously described simulation study results generalize to cases where there are more than two levels (groups) in a categorical moderator. To probe this idea, we included a third level in the categorical moderator and repeated the simulation studies examining the Type I error rate and statistical power. The results were qualitatively the same as those with two groups. In cases where the population had equal $\tau_i^2$ *and* had equal study sample sizes, the Type I error rate was relatively well controlled regardless of whether a MEM or MELSM was used. Again, the exceptions to this were cases where the groups had few studies. When there were imbalances in either $\tau_i^2$ or study sample sizes, then using a standard MEM led to inflated or conservative Type I error rates. Further, the power rates for

[5] Technically, the standard errors are for the corresponding $z$-test (Hedges & Pigott, 2005) and KH method (Knapp & Hartung, 2003), but this does not affect our arguments here given their equivalence with two groups (Section Tests for Categorical Moderators).

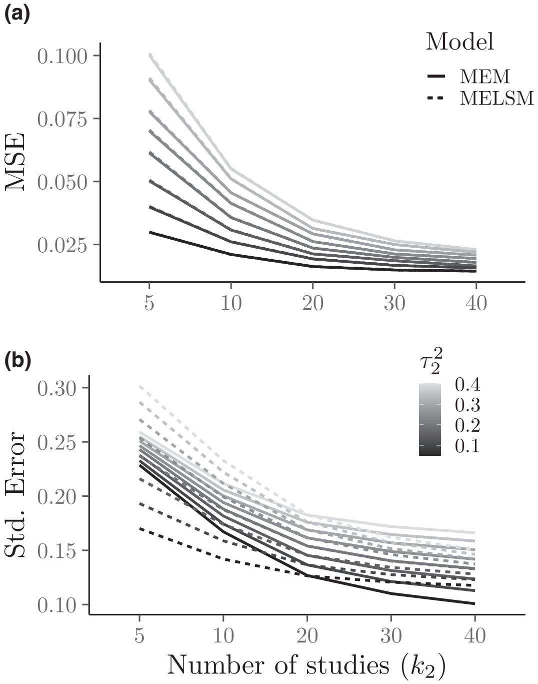

FIGURE 5 Panel a presents the MSE for the moderator effect ($\beta_1$) under the MEM and the MELSM, and panel b shows the corresponding average standard error. The grey lines denote $\tau_2^2$ values spanning from .05 to .4 in increments of .05 (dark to light). As can be clearly seen by the overlap of lines, there is no difference in estimation accuracy by assuming equal or unequal between-study variances. There are discrepancies in standard errors, however. Under imbalanced group sizes and unequal $\tau_i^2$, the estimated standard errors obtained with the MEM were either too large or too small, depending on the condition. Consequently, the standard errors were responsible for too high and too low type I error rates and power rates.

the MELSM were as good as or better than the MEM in all conditions where the latter had a valid (i.e., noninflated) Type I error rate. The complete results for these simulations can be found in the appendix. Note that, although not directly studied here, our simulation results for three groups qualitatively extend to testing multiple categorical moderators simultaneously (i.e., >3 groups). This is because tests of moderator effects directly depend, not on the number of moderator variables, but rather the number of studies $k$ and number of groups $p$ (Equations 13–16). The number of groups $p$ will of course increase with a higher number of moderators, thereby reducing power if $p$ becomes large relative to $k$.

## 4 | ILLUSTRATIVE EXAMPLE

In this section we use an empirical data set to demonstrate the MELSM and contrast it against the standard MEM in an applied setting. The data set was first used to synthesize the findings of $k = 46$ studies investigating the relationship between depression and specificity of future thinking (Gamble et al., 2019) and is available in the **psymetadata** R package (Rodriguez & Williams, 2022), that is, whether higher levels of depression are correlated with vagueness in future thoughts. The original data set contained 89 effect sizes (Pearson's or biserial correlations), but because some of these effect sizes were collected from the same study, they did not conform with the standard assumption of independent effect sizes in meta-regression. This issue was resolved by averaging all effect sizes within a study,[6] a procedure that is routinely applied in meta-analyses to alleviate the nonindependence of effect sizes (Tipton et al., 2019).

[6] Although a three-level MELSM would accommodate dependent effect sizes, such models are not yet available in standard meta-analytic software.

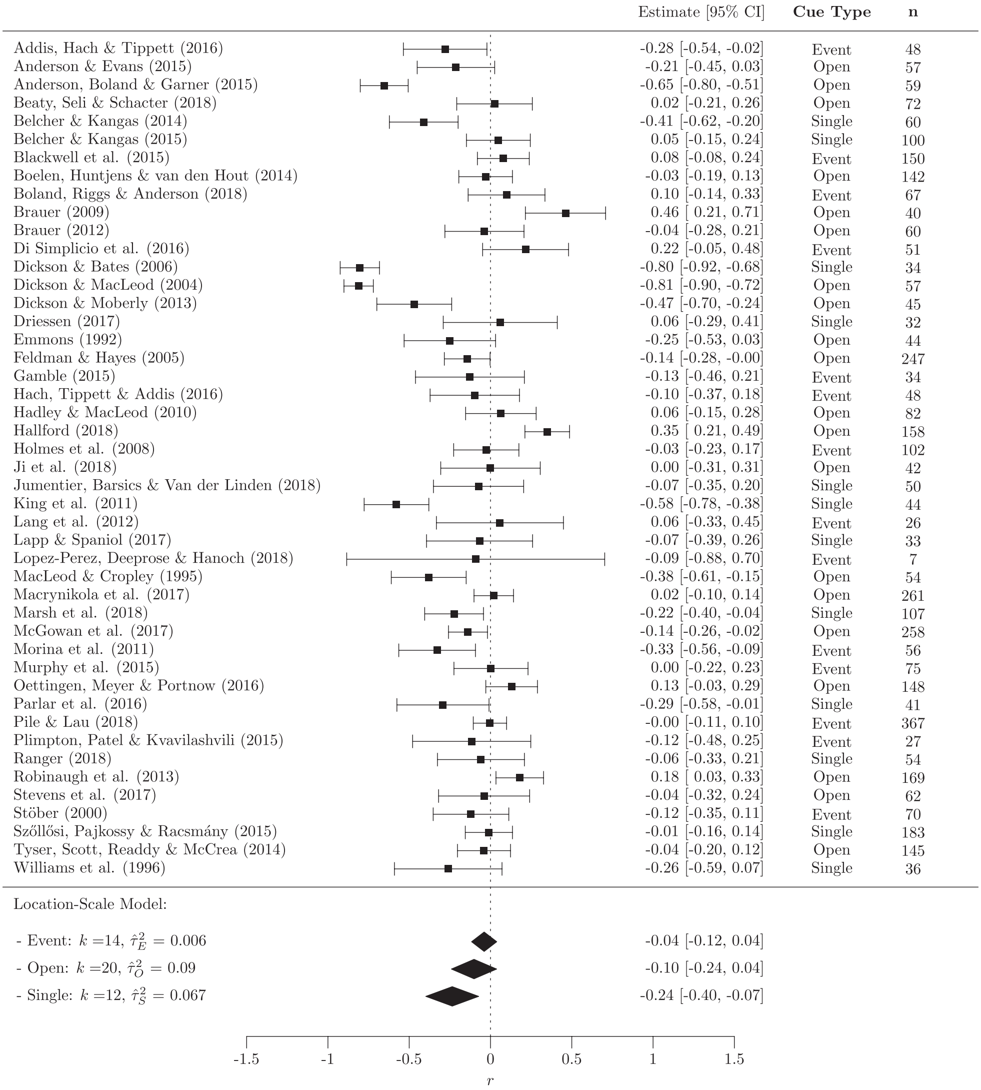

**FIGURE 6** Forest plot of effect sizes in Gamble et al. (2019) and results from MELSM with cue type as a moderator in location and scale.

One of the variables coded as a (trichotomous) moderator was the type of cue each study used to elicit participants' thoughts about a future event: a single word (e.g., “laughing”), event cues (e.g., “New Year's Eve”), or open (e.g., “I can imagine that, shortly, I…”). We thus fitted a MEM with cue type as a moderator for the location in addition to a MELSM with cue type as a moderator for both location and scale. The top of Figure 6 shows a forest plot (Lewis & Clarke, 2001) of the individual effect sizes, and the bottom displays the resulting moderator effect estimates from the MELSM.

Table 1 further shows, for the standard MEM and the MELSM, the estimates of the regression coefficients, their standard errors, and the resulting $p$-values. The KH method (equivalent to the $F$-test) was used to test the location coefficients in both models because it better maintains the nominal Type I error rate and valid power rates relative to the $z$-test (equivalent to the $Q$-test). Because the type of cue had three categories, it was coded as two dummy variables. Accordingly, the intercept $\beta_0$ captures the effect for the event cue type, $\beta_1$ captures the difference between the event cue type and the open cue type, and $\beta_2$ captures the difference between the event cue type and the single word cue type. As in the simulation

**TABLE 1** Regression estimates for example data

| | MEM | | | MELSM | | |
|---|---|---|---|---|---|---|
| | Est. | SE | $p$ | Est. | SE | $p$ |
| $\beta_0$ | −.05 | .07 | .52 | −.04 | .04 | .32 |
| $\beta_1$ | −.05 | .09 | .57 | −.06 | .08 | .43 |
| $\beta_2$ | −.19 | .11 | .09 | −.20 | .09 | .03 |

*Note*: Estimates of location coefficients under the standard MEM and the MELSM.

studies, the location estimates were similar between the two models, but the MELSM had smaller standard errors for all three coefficients. Consequently, all *p*-values for the moderator effects were smaller under the MELSM, and, of particular note, $\beta_2$ was not statistically significant for the MEM at the $a = .05$ level, whereas it was for the MELSM. The discrepancy in *p*-values between the two models reinforces the simulation study findings that inferences drawn about the location coefficients may differ depending on whether a MEM or a MELSM is fit to the data. The differences in standard errors were due to the differences in the heterogeneity estimates between the models. The (pooled) estimate for heterogeneity under the MEM was $\hat{\tau}^2 = .06$ but varied across groups for the MELSM. For the event, open, and single cue types, the heterogeneity estimates were $\hat{\tau}_E^2 = .006$, $\hat{\tau}_O^2 = .09$, and $\hat{\tau}_S^2 = .067$, respectively. That is, under the MELSM, the groups were estimated to have .1, 1.5, and 1.1 as much heterogeneity as they would have been estimated to have under the MEM.

In applied settings, a commonly estimated metric in mixed-effects meta-regressions is pseudo-$R^2$—a measure of how much heterogeneity is explained by the moderators (López-López et al., 2014; Raudenbush, 1994). Under the standard MEM, this metric is calculated as

$$R^2 = 1 - \frac{\hat{\tau}^2_{\text{full}}}{\hat{\tau}^2_{\text{null}}}, \tag{23}$$

where $\hat{\tau}^2_{\text{null}}$ corresponds to the heterogeneity parameter obtained via a null model (i.e., one without predictors) and $\hat{\tau}^2_{\text{full}}$ corresponds to the heterogeneity parameter obtained via the model containing moderators. Because the MELSM instead yields $\tau_i^2$, a separate $R^2$ can be calculated for each group

$$R_i^2 = 1 - \frac{\hat{\tau}^2_{i,\text{full}}}{\hat{\tau}^2_{i,\text{null}}}, \tag{24}$$

where $\hat{\tau}^2_{i,\text{null}}$ is the heterogeneity parameter estimate for the *i*th group under a null location submodel. That is, the $\hat{\tau}^2_{i,\text{null}}$ estimates are obtained by fitting a MELSM with the categorical moderator in the scale component, but not the location. Further, $\hat{\tau}^2_{i,\text{full}}$ is obtained by fitting the MELSM with the moderator in both the location and scale components of the model. Having separate $R_i^2$'s enables researchers to assess the differential explanatory power of a moderator.

In this example, pseudo-$R^2$ was estimated to be 0 under the MEM, indicating that cue type had no explanatory power for the heterogeneity in effect sizes. Under the MELSM, the pseudo-$R_i^2$ metrics were found to be $R_{\text{E}}^2 = .29$, $R_{\text{O}}^2 = 0$, and $R_{\text{S}}^2 = .22$ for the event, open, and single cue types, respectively. In other words, the results under the MELSM indicate that the event and single cue types explain more than 20% of the heterogeneity in their respective groups. Thus, one might conclude that the heterogeneity in the correlation between depression and the specificity of future thought may be well explained by event and single cue types, but not in the open cue type. We refer interested readers to Williams et al. (2021) for more information on computing and interpreting pseudo-$R_i^2$ in MELSMs.

Lastly, because the MEM was nested within the MELSM, we compared the two models using a likelihood ratio test. Recall that the MEM can be recast as a MELSM with only an intercept included in the scale component. Because the models did not differ in terms of their location coefficients, they did not need to be refit using maximum likelihood instead of REML. The likelihood ratio test indicated that the MELSM provided an improved fit over the MEM $\chi^2(2) = 7.8, p = .02$.

## 5 | DISCUSSION

In this work we set out to establish that in testing moderator effects with a mixed-effects meta-regression, researchers should assume unequal between-study variances by default. We illustrated how meta-analytic moderator tests were analogous to the *t*-test and ANOVA, so assuming unequal variances in moderator tests is no different than assuming unequal variances in a classical test of mean differences—the latter of which is already common practice in many disciplines.

As evidenced by two simulation studies, there are few costs in terms of Type I error (or statistical power) by assuming unequal between-study variances and potentially serious drawbacks to assuming them equal. When the population between-study variances differ between groups in a categorical moderator, statistical tests of moderator effects using the standard MEM can result in either grossly inflated or overly conservative Type I error rates. In turn, these error rates result in power rates that are too low or misleadingly high. Meanwhile, when moderator tests are carried out using the MELSM, the Type I error is well controlled regardless of equal variances or balanced sample sizes (except when there are few studies) and maintains valid power rates. These results substantiate the notion that researchers should assume unequal between-study variances as a default strategy.

As demonstrated in the illustrative example, MELSMs can be used to differentially assess the explanatory power of a categorical moderator. This is possible because the pseudo-$R_i^2$ statistic is based on the group-specific heterogeneity estimates, $\tau_i^2$. Consistent with this idea, other statistics that are traditionally based on $\tau^2$ can be obtained at the group level by instead using $\tau_i^2$. For instance, Williams et al. (2021) used a MELSM to compute country-specific $I^2$ (Higgins & Thompson, 2002) values in a meta-analysis.

### 5.1 | Future directions and limitations

This work focused on categorical moderators owing to their popularity in meta-analysis. However, a benefit of meta-regression techniques is the ability to accommodate both categorical and continuous predictors. Accordingly, it is possible to include a continuous variable as a moderator for the between-study variance in a MELSM. Indeed, it has been suggested that effect size heterogeneity may be inversely proportional to study sample size (Bowater & Escarela, 2013). Given that there is currently a dearth of studies on continuous moderators of between-study heterogeneity, investigating the extent to which a MELSM can be considered by default with a continuous moderator constitutes an interesting avenue for future exploration.

We demonstrated that the MELSM generally does well in controlling the Type I error rate, but we did so based on the implicit assumption that the included moderator was fixed in advance (e.g., through preregistration of hypotheses). That is, the observed error rates hold only when the moderator being tested is chosen a priori and not as a result of a data-driven process (see e.g., Berk et al., 2013). When the final meta-analytic model is chosen as a result of which moderators are (non-)significant, a potential consequence is an inflated Type I error rate. Although model selection techniques are relatively common in practice (Tipton et al., 2019), their use may contribute to recent concerns on the reproducibility of meta-analyses (Lakens et al., 2016).

In recent years, meta-analytic recommendations have demonstrated a preference for methods that reflect complex dependencies among effect sizes (e.g., multiple effect sizes nested within a study). To deal with such complexities in the variance structure, researchers can employ multivariate meta-analysis (Kalaian & Raudenbush, 1996; Raudenbush et al., 1988), robust variance estimation techniques (Hedges et al., 2010; Tipton, 2015), or three-level random-effects models (Konstantopoulos, 2011; Van den Noortgate et al., 2013). The MELSM described in this work is a two-level model and allows only a single effect size per study. It is thus limited in reflecting the dependencies among effect sizes. However, we note that averaging effect sizes within studies or randomly selecting a single effect size per study remains the dominant form of eliminating within-study dependencies (Tipton et al., 2019). In these cases, the MELSM can still be fruitfully applied. Future research should aim to understand whether the arguments made throughout this paper similarly apply to three-level MELSMs.

Finally, it has been recommended that at least five studies should be included per group when testing a categorical moderator in order to obtain trustworthy estimates of $\tau_i^2$ (Borenstein et al., 2009). Others have recommended that when the total number of studies is less than 20, the $\hat{\tau}_i^2$ should be pooled in order to more accurately estimate the between-study heterogeneity (Rubio-Aparicio et al., 2020). We agree with this suggestion when the focal interest lies in quantifying the between-study heterogeneity, but, as suggested by our simulation studies, when the focal interest is in testing the effect of a moderator, assuming equal $\tau_i^2$ when they are truly unequal and when there is an imbalance in study sample size may still lead to inadequate error rates.

## 5.2 | Conclusion

Presently, MELSMs can be easily fit in the R statistical computing environment with the R packages **metafor** (Viechtbauer, 2021) and **blsmeta** (Williams et al., 2021). Our hope is that this work will serve as an impetus for the wide-scale adoption of location-scale models in meta-analytic software and by applied researchers for testing categorical moderators.

## AUTHOR CONTRIBUTIONS

**Josue E. Rodriguez:** Conceptualization; methodology; visualization; writing – original draft; writing – review and editing. **Donald R. Williams:** Conceptualization; supervision; writing – original draft; writing – review and editing. **Paul-Christian Bürkner:** Methodology; supervision; writing – original draft; writing – review and editing.

## FUNDING INFORMATION

All authors have read and approved the final version of this article. DRW was supported by a National Science Foundation Graduate Research Fellowship under Grant No. 1650042. P.-C.B was partially funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 - 390740016. We thank Brenton Wiernik for helpful comments and discussions on earlier versions of this draft.

## CONFLICT OF INTEREST

The authors have no conflicts of interest with respect to their authorship or the publication of this article.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available on the Open Science Framework at https://osf.io/gs6uw/.

## ORCID

*Josue E. Rodriguez* https://orcid.org/0000-0002-9092-4869
*Donald R. Williams* https://orcid.org/0000-0001-6735-8785
*Paul-Christian Bürkner* https://orcid.org/0000-0001-5765-8995

## REFERENCES

Bartlett, M. S. (1936). The information available in small samples. *Mathematical Proceedings of the Cambridge Philosophical Society*, *32*(4), 560–566. https://doi.org/10.1017/S0305004100019290

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, *41*(2), 802–837. https://doi.org/10.1214/12-AOS1077

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Borenstein, M., & Higgins, J. P. T. (2013). Meta-analysis and subgroups. *Prevention Science*, *14*(2), 134–143. https://doi.org/10.1007/s11121-013-0377-7

Bowater, R. J., & Escarela, G. (2013). Heterogeneity and study size in random-effects meta-analysis. *Journal of Applied Statistics*, *40*(1), 2–16. https://doi.org/10.1080/02664763.2012.700448

Brannick, M. T., Potter, S. M., Benitez, B., & Morris, S. B. (2019). Bias and precision of alternate estimators in meta-analysis: Benefits of blending Schmidt-hunter and Hedges approaches. *Organizational Research Methods*, *22*(2), 490–514. https://doi.org/10.1177/1094428117741966

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395. https://doi.org/10.32614/RJ-2018-017

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Dahlke, J. A., & Wiernik, B. M. (2020). Not restricted to selection research: Accounting for indirect range restriction in organizational research. *Organizational Research Methods*, *23*(4), 717–749. https://doi.org/10.1177/1094428119859398

DeGeest, D. S., & Schmidt, F. L. (2010). The impact of research synthesis methods on industrialorganizational psychology: The road from pessimism to optimism about cumulative knowledge. *Research Synthesis Methods*, *1*(3–4), 185–197. https://doi.org/10.1002/jrsm.22

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. *International Review of Social Psychology*, *30*(1), 92–101. https://doi.org/10.5334/irsp.82

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Gamble, B., Moreau, D., Tippett, L. J., & Addis, D. R. (2019). Specificity of future thinking in depression: A meta-analysis. *Perspectives on Psychological Science*, *14*(5), 816–834. https://doi.org/10.1177/1745691619851784

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hartung, J., Makambi, K. H., & Argaç, D. (2001). An extended ANOVA F-test with applications to the heterogeneity problem in meta-analysis. *Biometrical Journal*, *43*(2), 135–146. https://doi.org/10.1002/1521-4036(200105)43:2<135::AID-BIMJ135>3.0.CO;2-H

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, *64*(2), 627–634. https://doi.org/10.1111/j.1541-0420.2007.00924.x

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, *31*(27), 3328–3336. https://doi.org/10.1002/sim.5338

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics*, *17*(4), 279–296. https://doi.org/10.3102/10769986017004279

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

Hedges, L. V., & Pigott, T. (2005). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, *9*, 426–445. https://doi.org/10.1037/1082-989X.9.4.426

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*(3), 203–217.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Method*s, 3(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 137–159. https://doi.org/10.1111/j.1467-985X.2008.00552.x

Ioannidis, J. P. A., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*, *335*(7626), 914–916. https://doi.org/10.1136/bmj.39343.408449.80

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, *1*(3), 227–235. https://doi.org/10.1037/1082-989X.1.3.227

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710. https://doi.org/10.1002/sim.1482

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, *2*(1), 61–76. https://doi.org/10.1002/jrsm.35

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*(1), 24. https://doi.org/10.1186/s40359-016-0126-3

Langan, D., Higgins, J. P. T., & Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. *Research Synthesis Methods*, *8*(2), 181–198. https://doi.org/10.1002/jrsm.1198

Lee, Y., & Nelder, J. A. (2006). Double hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *55*(2), 139–185. https://doi.org/10.1111/j.1467-9876.2006.00538.x

Lewis, S., & Clarke, M. (2001). Forest plots: Trying to see the wood and the trees. *BMJ: British Medical Journal*, *322*(7300), 1479–1480.

López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W., & Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 30–48. https://doi.org/10.1111/bmsp.12002

Murphy, B. P. (1967). Some two-sample tests when the variances are unequal: A simulation study. *Biometrika*, *54*(3–4), 679–683. https://doi.org/10.1093/biomet/54.3-4.679

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554.

R Core Team. (2021). *R: A language and environment for statistical computing*. Manual. R Foundation for Statistical Computing.

Raudenbush, S. W. (1994). Random effects models. In *The handbook of research synthesis* (pp. 301–321). Russell Sage Foundation.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, *103*(1), 111–120. https://doi.org/10.1037/0033-2909.103.1.111

Rodriguez, J. E., & Williams, D. R. (2022). Psymetadata: An R package containing open datasets from meta-analyses in psychology. *Journal of Open Psychology Data*, *10*(1), 8. https://doi.org/10.5334/jopd.61

Rubio-Aparicio, M., López-López, J. A., Viechtbauer, W., Marín-Martínez, F., Botella, J., & Sánchez-Meca, J. (2020). Testing categorical moderators in mixed-effects meta-analysis in the presence of heteroscedasticity. *The Journal of Experimental Education*, *88*(2), 288–310. https://doi.org/10.1080/00220973.2018.1561404

Rubio-Aparicio, M., Sánchez-Meca, J., López-López, J. A., Botella, J., & Marín-Martínez, F. (2017). Analysis of categorical moderators in mixed-effects meta-analysis: Consequences of using pooled versus separate estimates of the residual between-studies variances. *British Journal of Mathematical and Statistical Psychology*, *70*(3), 439–456. https://doi.org/10.1111/bmsp.12092

Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods*, *13*(1), 31–48. https://doi.org/10.1037/1082-989X.13.1.31

Scheffé, H. (1959). *The analysis of variance*. John Wiley & Sons.

Schoemann, A. M. (2016). Using multiple group modeling to test moderators in meta-analysis. *Research Synthesis Methods*, *7*(4), 387–401. https://doi.org/10.1002/jrsm.1200

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine*, *18*(20), 2693–2708. https://doi.org/10.1002/(SICI)1097-0258(19991030)18:20<2693::AID-SIM235>3.0.CO;2-V

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393. https://doi.org/10.1037/met0000011

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, *10*(2), 180–194. https://doi.org/10.1002/jrsm.1339

van Agteren, J., Iasiello, M., Lo, L., Bartholomaeus, J., Kopsaftis, Z., Carey, M., & Kyrios, M. (2021). A systematic review and meta-analysis of psychological interventions to improve mental wellbeing. *Nature Human Behaviour*, *5*(5), 631–652. https://doi.org/10.1038/s41562-021-01093-w

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *psychological Bulletin*From 1990–2013. *Journal of Open Psychology Dat*a, 5(1), 4. https://doi.org/10.5334/jopd.33

Viechtbauer, W. (2010). Conducting meta-analyses in *R* with the metafor Package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W. (2021). Metafor: A meta-analysis package for R. Manual.

Viechtbauer, W., & López-López, J. A. (2022). Location-scale models for meta-analysis. *Research Synthesis Methods*, *13*(6), 697–715. https://doi.org/10.1002/jrsm.1562

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods*, *20*(3), 360–374. https://doi.org/10.1037/met0000023

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*(3/4), 350–362. https://doi.org/10.2307/2332010

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, *34*(1/2), 28. https://doi.org/10.2307/2332510

Williams, D. R., Rodriguez, J. E., & Bürkner, P.-C. (2021). Putting variation into variance: Modeling between-study heterogeneity in meta-analysis. *PsyArXiv*, 1–17. https://doi.org/10.31234/osf.io/9vkqy

**How to cite this article:** Rodriguez, J. E., Williams, D. R., & Bürkner, P.-C. (2022). Heterogeneous heterogeneity by default: Testing categorical moderators in mixed-effects meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *00*, 1–32. https://doi.org/10.1111/bmsp.12299

# APPENDIX A

## A.1 | COMPLETE SIMULATION RESULTS (TWO GROUPS)

Power for $\beta_1 = .5$ (Table A1)

TABLE A1 Power for $\beta_1 = .5$

| | | *Q*-test | | *F*-test | |
|---|---|---|---|---|---|
| $\tau_2^2$ | $k_2$ | MEM | MELSM | MEM | MELSM |
| .05 | 5 | .614 | .813 | .544 | .782 |
| | 10 | .867 | .926 | .839 | .908 |
| | 20 | .970 | .969 | .964 | .964 |
| | 30 | .989 | .980 | .986 | .976 |
| | 40 | .993 | .982 | .992 | .98 |
| .10 | 5 | .576 | .701 | .527 | .664 |
| | 10 | .828 | .864 | .794 | .842 |
| | 20 | .945 | .944 | .939 | .939 |
| | 30 | .978 | .971 | .972 | .962 |
| | 40 | .984 | .976 | .983 | .973 |
| .15 | 5 | .547 | .618 | .501 | .58 |
| | 10 | .780 | .799 | .754 | .775 |
| | 20 | .920 | .920 | .913 | .913 |
| | 30 | .957 | .950 | .956 | .951 |
| | 40 | .974 | .967 | .973 | .966 |
| .20 | 5 | .522 | .540 | .491 | .514 |
| | 10 | .741 | .739 | .712 | .713 |
| | 20 | .888 | .887 | .878 | .877 |
| | 30 | .937 | .937 | .926 | .924 |
| | 40 | .958 | .956 | .955 | .953 |
| .25 | 5 | .513 | .503 | .465 | .462 |
| | 10 | .705 | .688 | .683 | .661 |
| | 20 | .856 | .858 | .839 | .839 |
| | 30 | .913 | .920 | .895 | .902 |
| | 40 | .938 | .945 | .933 | .941 |

TABLE A1 (Continued)

|  |  | *Q*-test |  | *F*-test |  |
|---|---|---|---|---|---|
| $\tau_2^2$ | $k_2$ | MEM | MELSM | MEM | MELSM |
| .30 | 5 | .501 | .463 | .47 | .422 |
|  | 10 | .680 | .642 | .644 | .6 |
|  | 20 | .823 | .824 | .81 | .811 |
|  | 30 | .885 | .896 | .874 | .887 |
|  | 40 | .918 | .933 | .904 | .922 |
| .35 | 5 | .495 | .430 | .454 | .393 |
|  | 10 | .656 | .599 | .619 | .566 |
|  | 20 | .793 | .793 | .771 | .773 |
|  | 30 | .859 | .881 | .847 | .868 |
|  | 40 | .890 | .915 | .875 | .906 |
| .40 | 5 | .486 | .404 | .438 | .361 |
|  | 10 | .627 | .555 | .593 | .526 |
|  | 20 | .761 | .762 | .746 | .747 |
|  | 30 | .828 | .856 | .807 | .839 |
|  | 40 | .864 | .902 | .852 | .895 |

*Note*. "MEM" indicates power rates observed under the standard MEM and "MELSM" indicates power rates observed under the MELSM with the moderator included in the scale submodel.

### Power for $\beta_1 = .2$ (Table A2).

TABLE A2 Power for $\beta_1 = .2$

|  |  | *Q*-test |  | *F*-test |  |
|---|---|---|---|---|---|
| $\tau_2^2$ | $k_2$ | MEM | MELSM | MEM | MELSM |
| .05 | 5 | .09 | .25 | .07 | .22 |
|  | 10 | .21 | .30 | .18 | .28 |
|  | 20 | .35 | .35 | .33 | .33 |
|  | 30 | .45 | .38 | .43 | .37 |
|  | 40 | .50 | .39 | .50 | .39 |
| .10 | 5 | .11 | .21 | .09 | .19 |
|  | 10 | .20 | .26 | .17 | .22 |
|  | 20 | .31 | .31 | .29 | .29 |
|  | 30 | .38 | .35 | .37 | .33 |
|  | 40 | .43 | .37 | .41 | .35 |
| .15 | 5 | .13 | .19 | .11 | .17 |
|  | 10 | .21 | .23 | .17 | .20 |
|  | 20 | .28 | .28 | .25 | .25 |
|  | 30 | .33 | .32 | .31 | .30 |
|  | 40 | .37 | .35 | .34 | .32 |

*(Continues)*

TABLE A2 (Continued)

| | | *Q*-test | | *F*-test | |
|---|---|---|---|---|---|
| $\tau_2^2$ | $k_2$ | MEM | MELSM | MEM | MELSM |
| .20 | 5 | .15 | .18 | .12 | .16 |
| | 10 | .20 | .21 | .18 | .19 |
| | 20 | .25 | .25 | .24 | .24 |
| | 30 | .30 | .31 | .28 | .28 |
| | 40 | .32 | .32 | .31 | .32 |
| .25 | 5 | .15 | .17 | .14 | .15 |
| | 10 | .20 | .19 | .17 | .16 |
| | 20 | .24 | .24 | .21 | .21 |
| | 30 | .27 | .28 | .24 | .26 |
| | 40 | .28 | .31 | .27 | .29 |
| .30 | 5 | .16 | .16 | .14 | .15 |
| | 10 | .20 | .18 | .17 | .15 |
| | 20 | .22 | .22 | .20 | .20 |
| | 30 | .23 | .26 | .23 | .25 |
| | 40 | .26 | .29 | .24 | .28 |
| .35 | 5 | .18 | .17 | .15 | .14 |
| | 10 | .19 | .16 | .16 | .14 |
| | 20 | .21 | .21 | .19 | .19 |
| | 30 | .21 | .24 | .20 | .23 |
| | 40 | .23 | .28 | .21 | .26 |
| .40 | 5 | .19 | .15 | .16 | .14 |
| | 10 | .19 | .16 | .17 | .14 |
| | 20 | .21 | .21 | .18 | .18 |
| | 30 | .20 | .24 | .18 | .22 |
| | 40 | .21 | .27 | .19 | .25 |

*Note*: $k_1 = 20$; $\tau_1^2 = .2$. "MEM" indicates power rates observed under the standard MEM and "MELSM" indicates power rates observed under the MELSM with the moderator included in the scale submodel.

Power for $\beta_2 = .8$ (Table A3).

TABLE A3 Power for $\beta_1 = .8$

| | | *Q*-test | | *F*-test | |
|---|---|---|---|---|---|
| $\tau_2^2$ | $k_2$ | MEM | MELSM | MEM | MELSM |
| .05 | 5 | .97 | .99 | .95 | .99 |
| | 10 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 20 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 30 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| .10 | 5 | .94 | .97 | .93 | .96 |
| | 10 | 1.00 | 1.00 | .99 | 1.00 |
| | 20 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 30 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 40 | 1.00 | 1.00 | 1.00 | 1.00 |

**TABLE A3** (Continued)

| | | *Q*-test | | *F*-test | |
|---|---|---|---|---|---|
| $\tau_2^2$ | $k_2$ | MEM | MELSM | MEM | MELSM |
| .15 | 5 | .92 | .93 | .90 | .91 |
| | 10 | .99 | .99 | .99 | .99 |
| | 20 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 30 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| .20 | 5 | .89 | .88 | .87 | .85 |
| | 10 | .98 | .98 | .98 | .98 |
| | 20 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 30 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| .25 | 5 | .87 | .83 | .85 | .81 |
| | 10 | .97 | .97 | .97 | .96 |
| | 20 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 30 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| .30 | 5 | .85 | .80 | .83 | .77 |
| | 10 | .96 | .95 | .96 | .94 |
| | 20 | 1.00 | 1.00 | .99 | .99 |
| | 30 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| .35 | 5 | .82 | .75 | .80 | .71 |
| | 10 | .96 | .94 | .94 | .92 |
| | 20 | .99 | .99 | .99 | .99 |
| | 30 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| .40 | 5 | .81 | .71 | .78 | .67 |
| | 10 | .94 | .90 | .92 | .89 |
| | 20 | .99 | .99 | .99 | .99 |
| | 30 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 40 | 1.00 | 1.00 | 1.00 | 1.00 |

*Note*: $k_1$ = 20; $\tau_1^2 = .2$. "MEM" indicates power rates observed under the standard MEM and "MELSM" indicates power rates observed under the MELSM with the moderator included in the scale submodel.

# APPENDIX B

## B.1 | SIMULATION RESULTS FOR THREE GROUPS

To examine whether the results in Section Simulation Studies generalized to situations involving the comparison of more than two groups, we conducted supplemental simulation studies. These followed the same procedures described in Simulation Studies but included a third level in the categorical moderator. Thus, in addition to $k_2$ and $\tau_2^2$, we also systematically varied the between-study variance and study sample size for the third group, $k_3$ and $\tau_3^2$. These quantities were varied according the values in Table B1. Within the columns for $k_i$ and $\tau_i^2$, each row represents a condition. For example, in the first condition for study sample size, $k_2 = k_3 = 5$, and for the between-study variance, $\tau_2^2 = \tau_3^2 = .2$ in the first condition. In the end, the total number of conditions was 2 ($Q$- or $F$-test) × 2 (MEM or MELSM) × 5 ($k_i$) × 9 ($\tau_i^2$) = 180 conditions. In each condition, the second and third groups were compared to the first group, whose values remained fixed at $\tau_1^2 = .2$ and $k_1 = 20$.

To study how well the MELSM controlled the Type I error of the $Q$- and $F$-tests relative to the MEM when there were more than two groups, we set $\beta_0 = \beta_1 = \beta_2 = 0$, or, in terms of group means, $\mu_1 = \mu_2 = \mu_3 = 0$. Both moderator tests were conducted for $\beta_1$ and $\beta_2$. The $a$ level for each test was set to .05. The results for $\beta_1$ can be seen in Table B2, while those for $\beta_2$ are presented in Table B3.

To study the empirical statistical power rates of the moderator tests with respect to the two moderator tests, we kept everything the same as in the Type I error rate simulation, with the exception that we changed the values of the moderator. For the power simulation study, the first moderator effect was set to $\beta_1 = .2$ or $\mu_2 - \mu_1 = .2$, and the second moderator effect was set to $\beta_1 = .8$ or $\mu_2 - \mu_1 = .8$. The power rates for the moderator effects can be found in Tables B4 and B5, respectively.

The patterns observed for the Type I error and power rates for the MEM and the MELSM qualitatively mirror those in the section Simulation Studies and are briefly described in the Discussion.

**TABLE B1** Simulation values

| $k_i$ | | $\tau_i^2$ | |
|---|---|---|---|
| $k_2$ | $k_3$ | $\tau_2^2$ | $\tau_3^2$ |
| 5 | 5 | .2 | .2 |
| 10 | 10 | .1 | .2 |
| 20 | 20 | .2 | .1 |
| 40 | 40 | .1 | .1 |
| 10 | 40 | .1 | .4 |
| 40 | 10 | .4 | .1 |
| | | .4 | .4 |
| | | .2 | .4 |
| | | .4 | .2 |

*Note*: $k_1 = 20$, $\tau_1^2 = .2$.

**TABLE B2** Type I error rate for $\beta_2$

| | | | | $Q$-test | | $F$-test | |
|---|---|---|---|---|---|---|---|
| $\tau_2^2$ | $\tau_3^2$ | $k_2$ | $k_3$ | MEM | MELSM | MEM | MELSM |
| .2 | .2 | 5 | 5 | .05 | .09 | .06 | .09 |
| | | 10 | 10 | .05 | .06 | .06 | .07 |
| | | 20 | 20 | .05 | .05 | .05 | .05 |
| | | 40 | 40 | .05 | .05 | .05 | .05 |
| | | 10 | 40 | .05 | .06 | .06 | .06 |
| | | 40 | 10 | .05 | .05 | .05 | .05 |
| .1 | .2 | 5 | 5 | .02 | .08 | .03 | .08 |
| | | 10 | 10 | .03 | .06 | .03 | .07 |
| | | 20 | 20 | .04 | .05 | .04 | .05 |
| | | 40 | 40 | .05 | .06 | .05 | .06 |
| | | 10 | 40 | .02 | .06 | .03 | .06 |
| | | 40 | 10 | .07 | .06 | .07 | .06 |

**TABLE B2** (Continued)

| | | | | *Q*-test | | *F*-test | |
|---|---|---|---|---|---|---|---|
| $\tau_2^2$ | $\tau_3^2$ | $k_2$ | $k_3$ | **MEM** | **MELSM** | **MEM** | **MELSM** |
| .2 | .1 | 5 | 5 | .05 | .09 | .07 | .10 |
| | | 10 | 10 | .06 | .06 | .07 | .07 |
| | | 20 | 20 | .07 | .06 | .08 | .06 |
| | | 40 | 40 | .07 | .05 | .08 | .06 |
| | | 10 | 40 | .09 | .06 | .09 | .07 |
| | | 40 | 10 | .06 | .05 | .06 | .06 |
| .1 | .1 | 5 | 5 | .03 | .08 | .03 | .07 |
| | | 10 | 10 | .04 | .06 | .05 | .07 |
| | | 20 | 20 | .06 | .05 | .07 | .06 |
| | | 40 | 40 | .08 | .06 | .08 | .05 |
| | | 10 | 40 | .05 | .06 | .06 | .06 |
| | | 40 | 10 | .07 | .05 | .08 | .06 |
| .1 | .4 | 5 | 5 | .01 | .08 | .02 | .08 |
| | | 10 | 10 | .02 | .06 | .02 | .06 |
| | | 20 | 20 | .02 | .05 | .02 | .06 |
| | | 40 | 40 | .03 | .06 | .03 | .06 |
| | | 10 | 40 | .01 | .06 | .01 | .06 |
| | | 40 | 10 | .05 | .05 | .06 | .06 |
| .4 | .1 | 5 | 5 | .11 | .10 | .13 | .10 |
| | | 10 | 10 | .10 | .07 | .11 | .07 |
| | | 20 | 20 | .08 | .06 | .09 | .06 |
| | | 40 | 40 | .06 | .05 | .06 | .05 |
| | | 10 | 40 | .15 | .07 | .14 | .07 |
| | | 40 | 10 | .04 | .05 | .04 | .06 |
| .4 | .4 | 5 | 5 | .09 | .10 | .11 | .10 |
| | | 10 | 10 | .06 | .06 | .08 | .07 |
| | | 20 | 20 | .04 | .06 | .05 | .06 |
| | | 40 | 40 | .02 | .05 | .03 | .05 |
| | | 10 | 40 | .05 | .06 | .05 | .07 |
| | | 40 | 10 | .03 | .05 | .03 | .06 |
| .2 | .4 | 5 | 5 | .04 | .08 | .05 | .09 |
| | | 10 | 10 | .03 | .06 | .04 | .07 |
| | | 20 | 20 | .03 | .05 | .03 | .06 |
| | | 40 | 40 | .03 | .06 | .03 | .06 |
| | | 10 | 40 | .02 | .06 | .02 | .07 |
| | | 40 | 10 | .04 | .05 | .04 | .06 |
| .4 | .2 | 5 | 5 | .10 | .10 | .12 | .11 |
| | | 10 | 10 | .08 | .06 | .09 | .07 |
| | | 20 | 20 | .06 | .06 | .07 | .06 |
| | | 40 | 40 | .04 | .05 | .05 | .06 |
| | | 10 | 40 | .09 | .06 | .10 | .07 |
| | | 40 | 10 | .03 | .05 | .04 | .06 |

*Note*: $k_1 = 20$; $\tau_1^2 = .2$. "MEM" indicates power rates observed under the standard MEM and "MELSM" indicates power rates observed under the MELSM with the moderator included in the scale submodel.

**TABLE B3** Type I error rate for $\beta_2$

| | | | | *Q*-test | | *F*-test | |
|---|---|---|---|---|---|---|---|
| $\tau_2^2$ | $\tau_3^2$ | $k_2$ | $k_3$ | MEM | MELSM | MEM | MELSM |
| .2 | .2 | 5 | 5 | .05 | .09 | .06 | .09 |
| | | 10 | 10 | .05 | .06 | .06 | .07 |
| | | 20 | 20 | .05 | .05 | .06 | .06 |
| | | 40 | 40 | .05 | .05 | .05 | .05 |
| | | 10 | 40 | .05 | .05 | .05 | .06 |
| | | 40 | 10 | .05 | .06 | .06 | .07 |
| .1 | .2 | 5 | 5 | .06 | .09 | .07 | .10 |
| | | 10 | 10 | .06 | .06 | .07 | .07 |
| | | 20 | 20 | .06 | .05 | .07 | .06 |
| | | 40 | 40 | .07 | .05 | .08 | .06 |
| | | 10 | 40 | .05 | .05 | .06 | .06 |
| | | 40 | 10 | .09 | .06 | .09 | .06 |
| .2 | .1 | 5 | 5 | .02 | .08 | .03 | .08 |
| | | 10 | 10 | .03 | .06 | .04 | .07 |
| | | 20 | 20 | .04 | .05 | .04 | .06 |
| | | 40 | 40 | .05 | .06 | .05 | .06 |
| | | 10 | 40 | .07 | .06 | .07 | .06 |
| | | 40 | 10 | .03 | .06 | .03 | .06 |
| .1 | .1 | 5 | 5 | .03 | .08 | .03 | .08 |
| | | 10 | 10 | .04 | .06 | .05 | .06 |
| | | 20 | 20 | .06 | .05 | .07 | .06 |
| | | 40 | 40 | .08 | .06 | .09 | .06 |
| | | 10 | 40 | .07 | .06 | .08 | .06 |
| | | 40 | 10 | .06 | .06 | .06 | .07 |
| .1 | .4 | 5 | 5 | .11 | .09 | .13 | .10 |
| | | 10 | 10 | .10 | .07 | .10 | .06 |
| | | 20 | 20 | .08 | .05 | .09 | .06 |
| | | 40 | 40 | .06 | .05 | .07 | .06 |
| | | 10 | 40 | .04 | .05 | .04 | .06 |
| | | 40 | 10 | .14 | .06 | .14 | .07 |
| .4 | .1 | 5 | 5 | .01 | .08 | .02 | .08 |
| | | 10 | 10 | .02 | .06 | .02 | .07 |
| | | 20 | 20 | .02 | .06 | .02 | .06 |
| | | 40 | 40 | .02 | .06 | .02 | .06 |
| | | 10 | 40 | .05 | .06 | .05 | .06 |
| | | 40 | 10 | .01 | .06 | .01 | .06 |
| .4 | .4 | 5 | 5 | .08 | .09 | .10 | .10 |
| | | 10 | 10 | .06 | .06 | .07 | .07 |
| | | 20 | 20 | .04 | .05 | .05 | .06 |
| | | 40 | 40 | .02 | .05 | .03 | .06 |
| | | 10 | 40 | .03 | .05 | .03 | .06 |
| | | 40 | 10 | .04 | .06 | .05 | .07 |

**TABLE B3** (Continued)

| | | | | *Q*-test | | *F*-test | |
|---|---|---|---|---|---|---|---|
| $\tau_2^2$ | $\tau_3^2$ | $k_2$ | $k_3$ | MEM | MELSM | MEM | MELSM |
| .2 | .4 | 5 | 5 | .10 | .09 | .12 | .10 |
| | | 10 | 10 | .08 | .06 | .09 | .07 |
| | | 20 | 20 | .06 | .05 | .07 | .06 |
| | | 40 | 40 | .05 | .06 | .05 | .05 |
| | | 10 | 40 | .04 | .05 | .04 | .06 |
| | | 40 | 10 | .09 | .06 | .10 | .07 |
| .4 | .2 | 5 | 5 | .04 | .08 | .05 | .10 |
| | | 10 | 10 | .03 | .06 | .04 | .07 |
| | | 20 | 20 | .03 | .06 | .03 | .06 |
| | | 40 | 40 | .02 | .06 | .03 | .06 |
| | | 10 | 40 | .04 | .05 | .04 | .05 |
| | | 40 | 10 | .02 | .06 | .02 | .07 |

*Note*: $k_1 = 20$; $\tau_1^2 = .2$. "MEM" indicates power rates observed under the standard MEM and "MELSM" indicates power rates observed under the MELSM with the moderator included in the scale submodel.

**TABLE B4** Power for $\beta_1 = .2$

| | | | | *Q*-test | | *F*-test | |
|---|---|---|---|---|---|---|---|
| $\tau_2^2$ | $\tau_3^2$ | $k_2$ | $k_3$ | MEM | MELSM | MEM | MELSM |
| .2 | .2 | 5 | 5 | .13 | .17 | .14 | .18 |
| | | 10 | 10 | .17 | .19 | .19 | .20 |
| | | 20 | 20 | .24 | .25 | .26 | .27 |
| | | 40 | 40 | .31 | .32 | .32 | .33 |
| | | 10 | 40 | .18 | .20 | .19 | .21 |
| | | 40 | 10 | .30 | .31 | .32 | .33 |
| .1 | .2 | 5 | 5 | .09 | .21 | .11 | .20 |
| | | 10 | 10 | .16 | .23 | .18 | .25 |
| | | 20 | 20 | .26 | .29 | .28 | .31 |
| | | 40 | 40 | .36 | .36 | .37 | .37 |
| | | 10 | 40 | .15 | .24 | .17 | .26 |
| | | 40 | 10 | .40 | .37 | .41 | .36 |
| .2 | .1 | 5 | 5 | .13 | .17 | .16 | .18 |
| | | 10 | 10 | .21 | .19 | .22 | .20 |
| | | 20 | 20 | .29 | .25 | .29 | .25 |
| | | 40 | 40 | .37 | .32 | .38 | .32 |
| | | 10 | 40 | .24 | .19 | .26 | .20 |
| | | 40 | 10 | .32 | .32 | .34 | .32 |
| .1 | .1 | 5 | 5 | .10 | .20 | .12 | .21 |
| | | 10 | 10 | .20 | .24 | .22 | .26 |
| | | 20 | 20 | .32 | .30 | .35 | .32 |
| | | 40 | 40 | .44 | .36 | .45 | .37 |
| | | 10 | 40 | .23 | .24 | .24 | .25 |
| | | 40 | 10 | .42 | .36 | .43 | .37 |

(Continues)

TABLE B4 (Continued)

| | | | | ***Q*-test** | | ***F*-test** | |
|---|---|---|---|---|---|---|---|
| $\tau_2^2$ | $\tau_3^2$ | $k_2$ | $k_3$ | **MEM** | **MELSM** | **MEM** | **MELSM** |
| .1 | .4 | 5 | 5 | .07 | .20 | .09 | .22 |
| | | 10 | 10 | .11 | .24 | .13 | .25 |
| | | 20 | 20 | .19 | .31 | .19 | .31 |
| | | 40 | 40 | .25 | .37 | .26 | .37 |
| | | 10 | 40 | .07 | .24 | .07 | .25 |
| | | 40 | 10 | .35 | .36 | .37 | .37 |
| .4 | .1 | 5 | 5 | .18 | .14 | .20 | .16 |
| | | 10 | 10 | .20 | .15 | .22 | .16 |
| | | 20 | 20 | .24 | .19 | .25 | .20 |
| | | 40 | 40 | .28 | .26 | .28 | .26 |
| | | 10 | 40 | .27 | .15 | .28 | .16 |
| | | 40 | 10 | .21 | .25 | .23 | .27 |
| .4 | .4 | 5 | 5 | .14 | .14 | .17 | .16 |
| | | 10 | 10 | .15 | .15 | .17 | .16 |
| | | 20 | 20 | .16 | .19 | .16 | .19 |
| | | 40 | 40 | .17 | .25 | .18 | .27 |
| | | 10 | 40 | .12 | .15 | .13 | .15 |
| | | 40 | 10 | .19 | .26 | .20 | .27 |
| .2 | .4 | 5 | 5 | .10 | .17 | .12 | .18 |
| | | 10 | 10 | .13 | .19 | .15 | .20 |
| | | 20 | 20 | .17 | .25 | .19 | .26 |
| | | 40 | 40 | .21 | .31 | .23 | .33 |
| | | 10 | 40 | .09 | .20 | .10 | .20 |
| | | 40 | 10 | .27 | .32 | .28 | .32 |
| .4 | .2 | 5 | 5 | .17 | .14 | .19 | .16 |
| | | 10 | 10 | .19 | .15 | .20 | .16 |
| | | 20 | 20 | .21 | .20 | .22 | .20 |
| | | 40 | 40 | .23 | .26 | .25 | .26 |
| | | 10 | 40 | .20 | .15 | .21 | .15 |
| | | 40 | 10 | .20 | .25 | .21 | .25 |

*Note*: $k_1 = 20$; $\tau_1^2 = .2$. “MEM” indicates power rates observed under the standard MEM and “MELSM” indicates power rates observed under the MELSM with the moderator included in the scale submodel.

**TABLE B5** Power for $\beta_2 = .8$

| | | | | *Q*-test | | *F*-test | |
|---|---|---|---|---|---|---|---|
| $\tau_2^2$ | $\tau_3^2$ | $k_2$ | $k_3$ | MEM | MELSM | MEM | MELSM |
| .2 | .2 | 5 | 5 | .87 | .87 | .90 | .89 |
| | | 10 | 10 | .98 | .98 | .99 | .98 |
| | | 20 | 20 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 10 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 10 | .98 | .98 | .99 | .98 |
| .1 | .2 | 5 | 5.00 | .88 | .87 | .90 | .88 |
| | | 10 | 10 | .99 | .98 | .99 | .98 |
| | | 20 | 20 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 10 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 10 | .99 | .98 | .99 | .98 |
| .2 | .1 | 5 | 5.00 | .93 | .96 | .94 | .96 |
| | | 10 | 10 | .99 | 1.00 | 1.00 | 1.00 |
| | | 20 | 20 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 10 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 10 | 1.00 | 1.00 | 1.00 | 1.00 |
| .1 | .1 | 5 | 5 | .94 | .96 | .95 | .97 |
| | | 10 | 10 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 20 | 20 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 10 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 10 | 1.00 | 1.00 | 1.00 | 1.00 |
| .1 | .4 | 5 | 5 | .80 | .69 | .82 | .71 |
| | | 10 | 10 | .94 | .90 | .96 | .91 |
| | | 20 | 20 | .99 | .99 | .99 | .99 |
| | | 40 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 10 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 10 | .97 | .90 | .97 | .91 |
| .4 | .1 | 5 | 5 | .90 | .96 | .93 | .97 |
| | | 10 | 10 | .99 | 1.00 | .99 | 1.00 |
| | | 20 | 20 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 10 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 10 | .98 | 1.00 | .98 | 1.00 |

*(Continues)*

TABLE B5 (Continued)

| | | | | Q-test | | F-test | |
|---|---|---|---|---|---|---|---|
| $\tau_2^2$ | $\tau_3^2$ | $k_2$ | $k_3$ | MEM | MELSM | MEM | MELSM |
| .4 | .4 | 5 | 5 | .77 | .70 | .79 | .71 |
| | | 10 | 10 | .92 | .90 | .93 | .91 |
| | | 20 | 20 | .98 | .99 | .98 | .99 |
| | | 40 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 10 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 10 | .90 | .91 | .91 | .91 |
| .2 | .4 | 5 | 5 | .79 | .69 | .81 | .71 |
| | | 10 | 10 | .93 | .90 | .95 | .91 |
| | | 20 | 20 | .99 | .99 | .99 | .99 |
| | | 40 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 10 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 10 | .95 | .91 | .95 | .91 |
| .4 | .2 | 5 | 5 | .85 | .87 | .87 | .88 |
| | | 10 | 10 | .97 | .98 | .98 | .98 |
| | | 20 | 20 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 10 | 40 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | 40 | 10 | .96 | .98 | .96 | .98 |

*Note*: $k_1$ = 20; $\tau_1^2 = .2$.“MEM” indicates power rates observed under the standard MEM and “MELSM” indicates power rates observed under the MELSM with the moderator included in the scale submodel.