Behavior Research Methods (2023) 55:1942–1964
https://doi.org/10.3758/s13428-022-01841-4

r2mlm: An R package calculating R-squared measures for multilevel models

Mairead Shaw[1] · Jason D. Rights[2] · Sonya S. Sterba[3] · Jessica Kay Flake[1]

Accepted: 18 March 2022 / Published online: 7 July 2022

Abstract
Multilevel models are used ubiquitously in the social and behavioral sciences and effect sizes are critical for contextualizing results. A general framework of R-squared effect size measures for multilevel models has only recently been developed. Rights and Sterba (2019) distinguished each source of explained variance for each possible kind of outcome variance. Though researchers have long desired a comprehensive and coherent approach to computing R-squared measures for multilevel models, the use of this framework has a steep learning curve. The purpose of this tutorial is to introduce and demonstrate using a new R package – *r2mlm* – that automates the intensive computations involved in implementing the framework and provides accompanying graphics to visualize all multilevel R-squared measures together. We use accessible illustrations with open data and code to demonstrate how to use and interpret the R package output.

Keywords Multilevel models · R-squared · Effect sizes · r2mlm

Introduction

Multilevel models (MLMs) are widely used in the behavioral sciences (Hox, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 2011). These models allow researchers to analyze clustered data structures that result from sampling and research designs across many areas of psychology. For example, students can be clustered within schools, people clustered within groups or dyads, and measurements clustered within person. Multilevel models can be used to avoid violations of the assumption of independence of observations for statistical tests and also allow researchers to explore dependencies and ask questions about the effects of individual- and cluster-level predictors on a given outcome.

Effect sizes are necessary for contextualizing the magnitude of the results from all kinds of statistical models and accurately conveying the properties of a sample. As such, journals and associations advise or require that effect sizes be reported (Cumming, 2014; Kelley & Preacher, 2012; Pek & Flora, 2018; Psychonomic Society, 2012). Historically, MLMs lacked a comprehensive approach for creating R-squared effect size measures that represented each distinct source of explained variance for each possible kind of outcome variance. Rights and Sterba (2019) addressed this shortcoming by developing an integrative R-squared effect size framework that, for the first time, utilized a complete partitioning of variance for MLMs. This framework provides separate measures corresponding to each potential source of explained variance that could account for total, within-cluster, or between-cluster outcome variance. The framework subsumes and expands on pre-existing MLM R-squared measures (from Aguinis & Culpepper, 2015; Bryk & Raudenbush, 1992; Hox, 2010; Johnson, 2014; Kreft & de Leeuw, 1998; Nakagawa & Schielzeth, 2013; Raudenbush & Bryk, 2002; Snijders & Bosker, 2011; Vonesh & Chinchilli, 1997; Xu, 2003). Analytic relationships between previous measures were provided in derivations in appendices of Rights and Sterba (2019).

The aim of the current work is to develop accessible implementation options for applied researchers to incorporate this integrative framework of effect sizes from Rights and Sterba (2019) into their empirical work. Using this R-squared framework properly has a steep learning curve

✉ Mairead Shaw
mairead.shaw@mail.mcgill.ca

[1] Department of Psychology, McGill University, 2001 McGill College, 7th Floor, Montreal, QC H3A 1G1, Canada

[2] Department of Psychology, University of British Columbia, Vancouver, BC, Canada

[3] Department of Psychology and Human Development, Vanderbilt University, Nashville, TN, USA

because it requires a thorough understanding of MLMs to conceptualize, interrelate, and visualize all of the R-squared measures in the framework together as a set. Additionally, it requires understanding how and why certain measures change when new terms are added to the multilevel model. For a researcher accustomed to a one-size-fits-all R-squared measure for single-level regression analyses, this MLM R-squared framework is substantially more involved. The fact that "popular software does not provide easy access" (Edwards et al., 2008, p. 6150) to MLM R-squared measures has been a longstanding impediment to their widespread and successful use in practice (Bickel, 2007; Demidenko et al., 2012; Jaeger et al., 2017; Kramer, 2005).

In this tutorial, we reduce the slope of this learning curve in two ways. First, we overview the basics of MLMs and the framework detailed in Rights and Sterba (2019, 2020). Second, we introduce and demonstrate a new R package, *r2mlm* (Shaw et al., 2020), that automates calculating all R-squared effect size measures described in the framework and provides accompanying graphics to visualize all of these R-squared measures together as an interrelated set. We demonstrate using this R package with openly available, simulated data examples accompanied by step-by-step code, and provide substantive interpretations of the resulting output. Given that R-squared measures are covered in virtually every MLM course, workshop, and textbook, this tutorial will benefit MLM users across the social and behavioral sciences.

Learning objectives and prerequisite knowledge

The learning objectives for this tutorial are to (1) understand the integrative R-squared framework detailed in Rights and Sterba (2019), (2) learn how to interpret the R-squared values for all measures in the framework, and (3) understand how to use the *r2mlm* R package to automate R-squared effect size computation and visualization. While we will briefly review multilevel modelling theory prior to walking through the examples, this tutorial is intended for researchers who are already familiar with specifying and interpreting MLMs and who wish to calculate R-squared effect sizes for their models. A researcher is sufficiently familiar with MLMs if they know MLMs partition variance into level 1/within-cluster variance and level 2/between-cluster variance, know the difference between fixed and random effects, and have specified MLMs and interpreted the resulting output in empirical research. Researchers unfamiliar with these aspects of MLMs are directed to McCoach (2010) and McCoach and Adelson (2010) for accessible yet brief introductions to MLMs. For those interested in comprehensive texts we suggest Raudenbush and Bryk (2002) or Snijders and Bosker (2011).

Though this R-squared effect size framework can be utilized with any software, when presenting our R functions, we will assume models were fit in R using the *lme4* or *nlme* packages, so it may be preferable (but is not necessary) to have some experience with R and *lme4* or *nlme*. For those without experience with R, a plethora of teaching resources are available. We recommend the first section of Wickham and Grolemund (2016), which is available for free online at www.r4ds.had.co.nz. Many more resources are aggregated at bigbookofr.com (Baruffa, 2021). For those without experience with *lme4* or *nlme* who want a formal introduction to the packages, we suggest Finch et al. (2014), or the documentation for each package (Bates et al., 2015; Pinheiro et al., 2020). Researchers wishing to run MLMs in other software can still use the effect size framework within R by manually entering parameter estimates, which we will demonstrate later.

Next, we review multilevel modelling theory and effect sizes, explain the R-squared framework developed by Rights and Sterba (2019), and subsequently demonstrate our new software tools to streamline and automate the application of this framework.

Brief overview: Multilevel modeling

Imagine you wish to examine the effect of student motivation on math test scores. You gather data from middle school students, and intend to run a linear regression with motivation as a predictor and math test score as the outcome. Many traditional statistical methods assume independence of observations. That is, controlling for motivation, students' math test scores will not otherwise be "paired, dependent, correlated, or associated in any way" (Glass & Hopkins, 1996, p. 295). When the assumption of independence is violated, the standard error is underestimated, which inflates type I error rates. Given that students in the same classroom have the same teacher, it is reasonable to suspect that students in the same class may be more similar in their math test scores (because of their shared experiences with teaching style, teaching experience, etc.) than to students in different classes, beyond the similarity accounted for by motivation. That is, there may be some degree of interdependence between math test scores among students in the same class.

We refer to this kind of data structure as being nested or clustered. One option for modelling clustered data is a multilevel model. These models are also known as random effects models, mixed models, and hierarchical linear models, among other names. Throughout this tutorial, we will use the term multilevel model (MLM). Multilevel models allow distinguishing variance within a cluster (e.g., how math scores of students vary within the same class) from

variance between clusters (e.g., how average math scores vary between classes). Instead of just one (fixed) intercept and one (fixed) slope per level 1 predictor, multilevel models allow for *cluster-specific* (random) intercepts and (random) slopes that accommodate the similarity of observations within a cluster. Multilevel models allow the researcher to answer questions at both the individual level (e.g., how does a student's motivation affect math test scores?) and the cluster level (e.g., how does teaching experience affect math test scores?) and to determine to what extent a model explains within-cluster (e.g., within-classroom) and between-cluster (e.g., between-classroom) outcome variation.

The following general equation for an MLM reflects the variance partitioning into within and between variance:

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\gamma}^{w} + \boldsymbol{z}'_{j}\boldsymbol{\gamma}^{b} + \boldsymbol{w}'_{ij}\boldsymbol{u}_{j} + r_{ij}. \quad (1)$$

In this equation, the bolded lowercase letters represent vectors, which stand in for all of the specific instances of each type of variable. For example, you could have five level 1 predictors in your model; in the above equation, all five are contained in the vector $\boldsymbol{x}_{ij}$. The y_{ij} is the outcome for a given unit, i, nested within a given cluster, j. The $\boldsymbol{\gamma}$ values represent fixed effects, i.e., the across-cluster average regression coefficients: $\boldsymbol{\gamma}^{w}$ is a vector of the within (i.e., level 1) fixed effects; $\boldsymbol{\gamma}^{b}$ is a vector of the between (i.e., level 2) fixed effects. $\boldsymbol{x}_{ij}$ is a vector of the level 1 predictors, and $\boldsymbol{z}_{j}$ a vector of the level 2 predictors (including a 1, for the intercept). $\boldsymbol{w}_{ij}$ is a vector consisting of 1 (again, for the intercept) and all level 1 predictors that have random slopes. $\boldsymbol{u}_{j}$ is a vector of the level 2 residuals (i.e., the random intercept residual and each random slope residual for cluster j), reflecting cluster-specific deviations from the across-cluster average regression coefficients. The r_{ij} is the residual for a given unit, i; that is, r_{ij} is the deviation of the outcome score from its cluster-specific expected outcome score conditional on the predictors and random effects.

Applied to our example of student math test scores predicted by motivation and teaching experience, we can express the multilevel regression equation as:

$$math_{ij} = \gamma_{00} + \gamma_{01}teaching_{ij} + \gamma_{10}motivation_{ij} + U_{0j} + U_{1j}motivation_{ij} + r_{ij} \quad (2)$$

Here, student math test scores ($math_{ij}$) are predicted by the level 1 variable motivation ($\gamma_{10} * motivation_{ij}$) with a random slope ($U_{1j} * motivation_{ij}$) and the level 2 variable teaching experience ($\gamma_{01} * teaching_{ij}$); the model also includes the fixed component of the intercept (γ_{00}) as well as the random component (U_{0j}), and the level 1 residual (r_{ij}). Thus, this model accounts for the variability in intercepts and slopes across classrooms and can address questions about how predictors at both the student and classroom level relate to the outcome.

Brief overview: Effect sizes

Per Kelley and Preacher (2012), the term "effect size" encapsulates any quantitative reflection of the magnitude of some phenomenon, with reference to a specific research question. This includes a variety of statistics, describing various aspects of a model. For example, standard deviation can describe variability and Cohen's d can describe differences between group means. Effect sizes can be standardized (e.g., Cohen's d, expressed in standard deviation units) or unstandardized (e.g., an estimated mean difference, expressed in the units of the dependent variable) (Pek & Flora, 2018). Reporting effect size measures appropriate for a given research question is important for contextualizing the results by providing an indication of *practical* significance (i.e., "how meaningful is this effect?") beyond just statistical significance.

One popular effect size in traditional statistical frameworks is R-squared, a standardized effect size computed as the proportion of variance explained by a model (Wright, 1921). Generically, it can be represented as the ratio of the outcome variance explained by the model to the total outcome variance:

$$R^2 = \frac{explained\ variance}{total\ variance} \quad (3)$$

This yields an intuitive variance explained measure ranging from 0 to 1, with 0 indicating 0% explained and 1 indicating 100% explained.

As detailed by Rights and Sterba (2019), for MLMs, calculating the proportion of variance explained is complicated by the fact that there are multiple types of outcome variance (total vs. within-cluster vs. between-cluster), in contrast to single-level regression models which have only one type of outcome variance. Moreover, in MLMs there are multiple sources that could contribute to explained variance (e.g., predictors at different levels via their fixed and random components) in contrast to single-level regression models which have only one source of explained variance (predictors at that single-level via their fixed components). Some researchers developing MLM R-squared measures had provided a single measure (e.g., Snijders & Bosker 1999, 2011) and sought an omnibus "one-size-fits-all" measure, analogous to that in single-level regression (e.g., Orelien & Edwards, 2008). Others have suggested pairs of measures (e.g., Hox, 2010; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002), but they collapse across sources of explained variance, or examine only one kind of outcome variance, and can yield misleading or uninterpretable results (see Rights & Sterba, 2019, 2020 for a thorough review). For example, Johnson (2014), Nakagawa and Schielzeth (2013), and Snijders and Bosker (1994, 2011) all presented measures based on

partitioning of model-implied variance in MLMs but did not use a *full* partitioning of outcome variance. These measures did not consider (1) partitioning variance into each of within, between, and total variance, (2) partitioning explained total variance into contributions by level 1 predictors versus level 2 predictors via fixed effects, or (3) partitioning explained variance into contributions via random slope versus via random intercept variation. Ultimately, no single or small set of MLM R-squared measures can thoroughly distinguish the contribution of each distinct source of variance for each applicable kind of outcome variance.

Rights and Sterba (2019) overcame these limitations by developing a framework that provides a comprehensive suite of R-squared measures that yields a complete picture of the model's explanatory power and provides new measures while subsuming pre-existing measures (Aguinis & Culpepper, 2015; Bryk & Raudenbush, 1992; Hox, 2010; Johnson, 2014; Kreft & de Leeuw, 1998; Nakagawa & Schielzeth, 2013; Raudenbush & Bryk, 2002; Snijders & Bosker, 2011; Vonesh & Chinchilli, 1997; Xu, 2003). To increase the accessibility of this framework, we released an R package called *r2mlm* that takes an MLM as input and calculates the R-squared values according to Rights and Sterba's (2019) framework (Shaw et al., 2020). To help develop users' intuitions about the framework and their comfort using the R package, the remainder of this paper will overview Rights and Sterba's (2019) framework, then walk through calculating and interpreting R-squared values using open data and code.

An R-squared framework for multilevel models

As mentioned, calculating variance explained for an MLM is complicated by total variance being partitioned into within and between variances. The Rights and Sterba (2019) framework intuitively maps variance explained for MLMs by considering variance explained at each of these levels – within variance explained and between variance explained – as well as the total variance (i.e., sum of within and between variance) explained. Here, we introduce the framework in plain language to provide an accessible guide, which supplements the published technical work.

At the within level of the model, there are three possible sources of variance: the level 1 predictors via the fixed effects (shorthand: "*f*1"), the level 1 predictors via the random effects (shorthand: "*v*"), and the level 1 residuals (shorthand: *resid*). Hence, a within-cluster R-squared measure has the following form:

$$R^2_{within} = \frac{explained\ within\ variance}{\text{var}_{f1} + \text{var}_{v} + \text{var}_{resid}} \tag{4}$$

Where var_{f1} denotes variance explained by *f*1, var_{v} denotes variance explained by *v*, and var_{resid} denotes residual variance. You can then calculate two distinct effect sizes from this: within variance explained by level 1 predictors via fixed effects (termed $R^{2(f1)}_{within}$) and within variance explained by level 1 predictors via random effects (termed $R^{2(v)}_{within}$). Note that a given R-squared is described by two elements: a subscript and a superscript. The subscripts indicate at what level variance is being explained: "*within*" for within-cluster, "*between*" for between-cluster, and "*total*" for total. The superscripts indicate what potential sources of variance are contributing to variance explained: "*f*1" for level 1 predictors via fixed effects, "*f*2" for level 2 predictors via fixed effects, and so on. For example, at the within level, the R-squared measure for the level 1 predictors via fixed effects is represented as $R^{2(f1)}_{within}$.

$$R^{2(f1)}_{within} = \frac{\text{var}_{f1}}{\text{var}_{f1} + \text{var}_{v} + \text{var}_{resid}} \tag{5}$$

$$R^{2(v)}_{within} = \frac{\text{var}_{v}}{\text{var}_{f1} + \text{var}_{v} + \text{var}_{resid}} \tag{6}$$

You can consider each of these effect sizes alone or add the two to consider variance explained by level 1 predictors via fixed and random effects combined, yielding $R^{2(f1v)}_{within} = R^{2(f1)}_{within} + R^{2(v)}_{within}$.

Between variance is composed of the contribution of level 2 predictors via fixed effects (shorthand: "*f*2") and cluster-specific means via intercept variation (shorthand: "*m*"), yielding the following expression for a between-cluster R-squared measure:

$$R^2_{between} = \frac{explained\ between\ variance}{\text{var}_{f2} + \text{var}_{m}} \tag{7}$$

You can then calculate two possible R-squared effect sizes, quantifying the between variance explained by each of the two between-cluster sources, respectively:

$$R^{2(f2)}_{between} = \frac{\text{var}_{f2}}{\text{var}_{f2} + \text{var}_{m}} \tag{8}$$

$$R^{2(m)}_{between} = \frac{\text{var}_{m}}{\text{var}_{f2} + \text{var}_{m}} \tag{9}$$

Here, there is no utility in combining these measures, as by definition they will account for the entirety of the between variance and hence will sum to 1 every time.

Total variance then is the combination of within and between variance explained, and thus total R-squared measures take the following form:

$$R^2_{total} = \frac{explained\ total\ variance}{\text{var}_{f1} + \text{var}_{f2} + \text{var}_{v} + + \text{var}_{m} + \text{var}_{resid}} \quad (10)$$

There are four component effect sizes, each quantifying total variance explained by the following sources, respectively: level 1 predictors via fixed effects ("*f*1"), level 2 predictors via fixed effects ("*f*2"), level 1 predictors via random slope variation ("*v*"), and cluster-specific outcome means via intercept variation ("*m*"):

$$R^{2(f1)}_{total} = \frac{\text{var}_{f1}}{\text{var}_{f1} + \text{var}_{f2} + \text{var}_{v} + \text{var}_{m} + \text{var}_{resid}} \quad (11)$$

$$R^{2(f2)}_{total} = \frac{\text{var}_{f2}}{\text{var}_{f1} + \text{var}_{f2} + \text{var}_{v} + \text{var}_{m} + \text{var}_{resid}} \quad (12)$$

$$R^{2(v)}_{total} = \frac{\text{var}_{V}}{\text{var}_{f1} + \text{var}_{f2} + \text{var}_{v} + \text{var}_{m} + \text{var}_{resid}} \quad (13)$$

$$R^{2(m)}_{total} = \frac{\text{var}_{m}}{\text{var}_{f1} + \text{var}_{f2} + \text{var}_{v} + \text{var}_{m} + \text{var}_{resid}} \quad (14)$$

Rights and Sterba (2019) recommend considering how much variance is explained by each individual component for the most complete information, but researchers can additionally add proportions together to consider more general questions like "how much variance is explained by all predictors via fixed effects?" ($R^{2(f)}_{total} = R^{2(f1)}_{total} + R^{2(f2)}_{total}$). You can also consider other combinations of these component effect sizes, for instance, total variance explained by predictors at both levels via fixed effects and random slopes ($R^{2(fv)}_{total} = R^{2(f1)}_{total} + R^{2(f2)}_{total} + R^{2(v)}_{total}$), and total variance explained by all sources ($R^{2(fvm)}_{total} = R^{2(f1)}_{total} + R^{2(f2)}_{total} + R^{2(v)}_{total} + R^{2(m)}_{total}$). The level 1 residuals are the remaining unexplained variance, so there is no component effect size "variance explained by unexplained variance."

Researchers may not be accustomed to considering random effect variation as "explained variance," which is the case with all aforementioned measures containing a v or m in the superscript (e.g., $R^{2(v)}_{within}$, $R^{2(m)}_{total}$). Previous MLM literature has offered two perspectives on how to treat variance attributable to random intercepts and slopes, called the "marginal" and "conditional" approaches (e.g., Edwards et al., 2008; Orelien & Edwards, 2008; Vonesh & Chinchilli, 1997; Wang & Schaalje, 2009; Xu, 2003). In the marginal approach, all variance attributable to predictors via random slope variation and attributable to cluster means via random intercept variation (i.e., sources "v" and "m") is treated as unexplained. In the conditional approach, variance attributable to predictors via random slope variation ("v") and/or attributable to cluster means via random intercept variation ("m") is treated as explained. Substantive justification for why one might want to consider a conditional R-squared measure was provided in Vonesh and Chinchilli (1997) and Rights and Sterba (2019).

The Rights and Sterba (2019) framework offers researchers access to both the marginal and conditional approaches, because it separately quantifies variance attributable to each source that would be entered into the numerator of either a marginal or conditional measure. The marginal approach is more common in psychology, whereas the conditional approach has received more attention in biostatistics (e.g., Vonesh & Chinchilli, 1997). Nonetheless, the conditional approach has actually been used for years in the social sciences without much recognition. For example, one of Raudenbush & Bryk's (1992, Raudenbush & Bryk, 2002) measures is actually a conditional measure. More broadly, the conditional approach may be useful for social science researchers to consider for descriptive purposes to quantify the degree of each kind of between-cluster heterogeneity. Otherwise the extent of such heterogeneity is often not discussed or is interpreted only qualitatively. For example, once a researcher realizes they have a large portion of variation attributable to predictors via random slope variation ($R^{2(v)}_{total}$), this could, in turn, motivate researchers to consider possible cross-level interaction terms in future modelling (Aguinis & Culpepper, 2015; Rights & Sterba, 2019, 2020). Relatedly, quantifying the extent of between-cluster outcome variance attributable to intercept variation ($R^{2(m)}_{total}$) can easily indicate to the researcher whether there are substantial differences between clusters beyond that explained by predictors. In psychology, random effect variation is often thought of as residual variance, so the idea of "residual variance" as "explained variance" can be unintuitive. A researcher wishing to quantify variation in intercepts and/or slopes (i.e., source "m" and/or "v") without thinking of it as "variance explained" can instead interpret it with the more neutral language of variance "attributable to" or "modeled by" the source(s).

Overall, the single-source R-squared measures defined in Equations 4–14, as well as the combinations described above, yield 12 different R-squared measures for a given model, as summarized in Table 1.

Framework assumptions

A few assumptions underlie this framework as originally delineated by Rights and Sterba (2019). This framework is implementable for the most common multilevel

Table 1 Definitions of multilevel model R^2 measures in integrative framework

Measure	Definition/Interpretation
Total MLM R^2 measures	
$R^{2(f1)}_{total} = \frac{var_{f1}}{var_{f1}+var_{f2}+var_{v}+var_{m}+var_{resid}}$	Proportion of total outcome variance explained by *level 1 predictors via fixed slopes*
$R^{2(f2)}_{total} = \frac{var_{f2}}{var_{f1}+var_{f2}+var_{v}+var_{m}+var_{resid}}$	Proportion of total outcome variance explained by *level 2 predictors via fixed slopes*
$R^{2(f)}_{total} = \frac{var_{f1}+var_{f2}}{var_{f1}+var_{f2}+var_{v}+var_{m}+var_{resid}}$	Proportion of total outcome variance explained by *all predictors via fixed slopes*
$R^{2(v)}_{total} = \frac{var_{v}}{var_{f1}+var_{f2}+var_{v}+var_{m}+var_{resid}}$	Proportion of total outcome variance explained by *level 1 predictors via random slope variation/covariation*
$R^{2(m)}_{total} = \frac{var_{m}}{var_{f1}+var_{f2}+var_{v}+var_{m}+var_{resid}}$	Proportion of total outcome variance explained by *cluster-specific outcome means via random intercept variation*
$R^{2(fv)}_{total} = \frac{var_{f1}+var_{f2}+var_{v}}{var_{f1}+var_{f2}+var_{v}+var_{m}+var_{resid}}$	Proportion of total outcome variance explained by *predictors via fixed slopes and random slope variation/covariation*
$R^{2(fvm)}_{total} = \frac{var_{f1}+var_{f2}+var_{v}+var_{m}}{var_{f1}+var_{f2}+var_{v}+var_{m}+var_{resid}}$	Proportion of total outcome variance explained by *predictors via fixed slopes and random slope variation/covariation* and by *cluster-specific outcome means via random intercept variation*
Within-cluster MLM R^2 measures	
$R^{2(f1)}_{within} = \frac{var_{f1}}{var_{f1}+var_{v}+var_{resid}}$	Proportion of within-cluster outcome variance explained by *level 1 predictors via fixed slopes*
$R^{2(v)}_{within} = \frac{var_{v}}{var_{f1}+var_{v}+var_{resid}}$	Proportion of within-cluster outcome variance explained by *level 1 predictors via random slope variation/covariation*
$R^{2(f1v)}_{within} = \frac{var_{f1}+var_{v}}{var_{f1}+var_{v}+var_{resid}}$	Proportion of within-cluster outcome variance explained by *level 1 predictors via fixed slopes and random slope variation/covariation*
Between-cluster MLM R^2 measures	
$R^{2(f2)}_{between} = \frac{var_{f2}}{var_{f2}+var_{m}}$	Proportion of between-cluster outcome variance explained by *level 2 predictors via fixed slopes*
$R^{2(m)}_{between} = \frac{var_{m}}{var_{f2}+var_{m}}$	Proportion of between-cluster outcome variance explained by *cluster-specific outcome means via random intercept variation*

A given R-squared is described by two elements: a subscript and a superscript. The subscripts indicate at what level variance is being explained: "*within*" for within-cluster, "*between*" for between-cluster, and "*total*" for total. The superscripts indicate what potential sources of variance are contributing to variance explained: "*f*1" for level 1 predictors via fixed effects, "*f*2" for level 2 predictors via fixed effects, "*v*" for level 1 predictors via random slope variation/covariation, "*m*" for cluster-specific outcome means via random intercept variation. Adapted from "Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures," by J. Rights and Sterba, 2019, *Psychological Methods*, *24*(3), p. 7. Copyright 2019 by the American Psychological Association.

specification: two-level multilevel models with normally distributed outcomes and homoscedastic residual variances. Initially in Rights and Sterba (2019), the framework assumed level 1 predictors were cluster-mean-centered, which avoids the pitfall of estimating conflated effects that are uninterpretable blends of level-specific effects (Enders & Tofighi, 2007; LaHuis et al., 2014; Raudenbush & Bryk, 2002). Subsequently, the full decomposition of variance was derived without assuming cluster-mean-centering of level 1 predictors (Rights & Sterba, 2021). Hence all total, within-cluster, and between-cluster R-squared measures in the framework are available for non-cluster-mean-centered models as well (Rights & Sterba, 2021), as we demonstrate later in this tutorial. In the Discussion, we also mention recent generalizations of this framework to accommodate additional modeling complexities, including heteroscedastic residual variance and alternative centering options, but here focus pedagogically on the original framework and

assumptions from Rights and Sterba (2019) due to its greater simplicity and widespread applicability.

R package

Broadly, this R-squared framework for multilevel models disaggregates each potential source of variance explained into distinct effect sizes at within, between, and total levels of the model. This allows comprehensive consideration of how each individual and/or composite term in the model contributes to the proportion of variance explained. The newly developed package *r2mlm* introduced in this tutorial paper facilitates calculating effect sizes with this underlying framework. To help develop readers' intuitions about the framework and illustrate using the R package, we will now demonstrate calculating and interpreting effect sizes for a variety of multilevel models using *r2mlm* in the context of accessible empirical examples.

Data demonstrations

Example data

For this tutorial, we will use simulated data included with the *r2mlm* package. To access the dataset and perform all analyses, the first step is to install and load the package.

```
install.packages("r2mlm")
library(r2mlm)
```

The simulated dataset included with the package is called *teachsat*, and contains information related to teacher job satisfaction. Teachers are clustered within schools, 30 teachers per school for 300 schools, for a total of 9000 observations. The dataset contains the following variables:

- *schoolID*: the school identification number, range from 1–300. This is our clustering variable.
- *teacherID*: a teacher's ID number within a school, range from 1–30
- *satisfaction*: teacher job satisfaction on a 1–10 scale (1 = low satisfaction)
- *control_c*: school-mean-centered teacher self-reported control over the curriculum (lower = less control)
- *control_m*: school mean rating of teacher's self-reported control over the curriculum
- *salary_c*: school-mean-centered teacher salary (thousands of dollars)
- *salary_m*: school mean teacher salary (thousands of dollars)
- *s_t_ratio*: student-teacher ratio (number of students per teacher)

For our examples, we will specify a variety of models predicting teacher job satisfaction. Throughout the examples we will evaluate the meaning of various effects through two lenses: standardized R-squared effect sizes, and unstandardized regression coefficients.

Null model

The null model contains only terms for the fixed and random components of the intercept of teacher job satisfaction. As such, the null model is also called the random-intercept-only model. It is usually the first model estimated because researchers can easily calculate the intraclass correlation coefficient (ICC) from it.

Level 1: $satisfaction_{ij} = \beta_{0j} + R_{ij}$
Level 2: $\beta_{0j} = \gamma_{00} + U_{0j}$
Combined: $satisfaction_{ij} = \gamma_{00} + U_{0j} + R_{ij}$

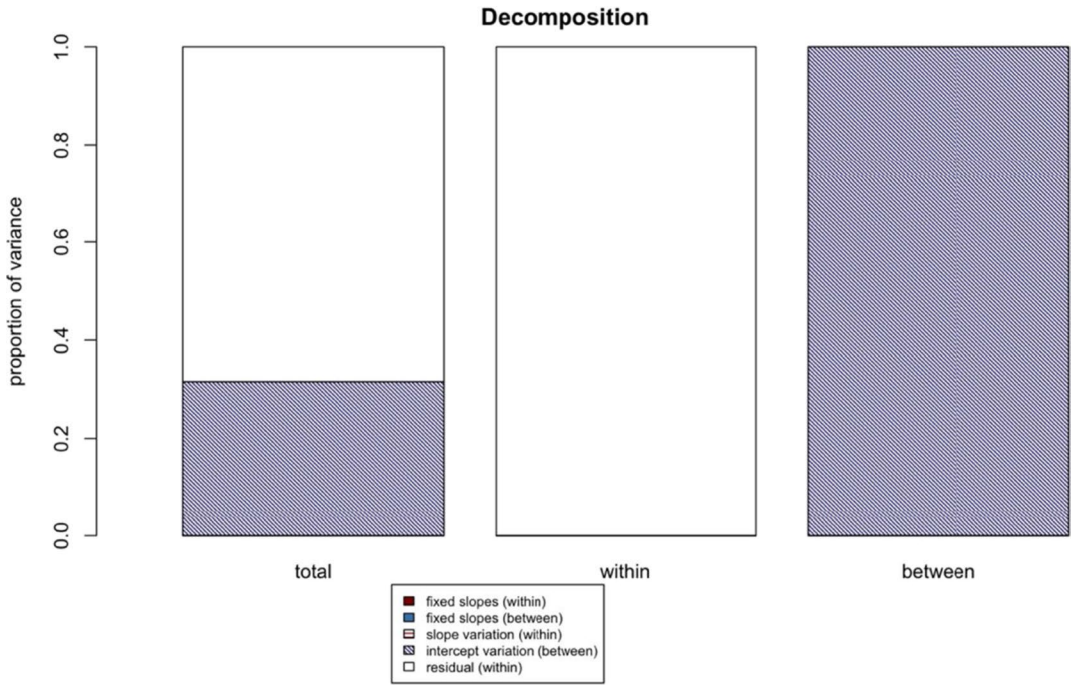
```
null_model <- lmer(satisfaction ~ 1 + (1 | schoolID),
                   data = teachsat,
                   REML = TRUE)

summary(null_model)

## Linear mixed model fit by REML ['lmerMod']
## Formula: satisfaction ~ 1 + (1 | schoolID)
##    Data: teachsat
##
## REML criterion at convergence: 30098.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8269 -0.6385  0.0012  0.6435  3.2874
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  schoolID (Intercept) 0.699    0.836
##  Residual             1.516    1.231
## Number of obs: 9000, groups:  schoolID, 300
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  5.99677    0.04998     120
```

This model produces one fixed effect estimate for the intercept. The predicted value of teacher satisfaction across all teachers in all schools, i.e., the predicted grand mean of satisfaction, is 6.00. To calculate effect sizes for a given model, we call *r2mlm(model_name)*. Note that *r2mlm* can handle models run using both *lme4* and *nlme*. For brevity, we demonstrate coding models using *lme4*, but the *r2mlm(model_name)* function call for calculating effect sizes for models is identical for those created using *nlme*.

```
r2mlm(null_model)
```




```
## $Decompositions
##                 total             within between
## fixed, within   0                 0      NA
## fixed, between  0                 NA     0
## slope variation 0                 0      NA
## mean variation  0.315546785367943 NA     1
## sigma2          0.684453214632058 1      NA
##
## $R2s
##     total             within between
## f1  0                 0      NA
## f2  0                 NA     0
## v   0                 0      NA
## m   0.315546785367943 NA     1
## f   0                 NA     NA
## fv  0                 0      NA
## fvm 0.315546785367943 NA     NA
```

There are three components to the function output. First, there is a bar chart that depicts the R-squared values. Second, there are variance decompositions. Third, there are the R-squared values specified in Rights and Sterba's (2019) framework and summarized in Table 1. Note that you can suppress the bar chart output with the *bargraph* argument: *r2mlm(model_name, bargraph = FALSE)*.

For the null model, intercept variation across schools (i.e., clusters) is the only thing accounting for variance in teacher job satisfaction. The function output aligns with our expectations: in this model, the total variance can only be explained with information we have about how school means vary on the outcome. Per the output, 31.6% of the total variance is accounted for by cluster membership, shown as "mean variation" in the decomposition output, as "m" in the R-squared output, and as "intercept variation (between)" in the total bar graph. Note that "fvm" in the R-squared output (i.e., $R^{2(fvm)}_{total}$) is a combination of variance attributable to predictors at both levels via fixed effects ("f"), to level 1 predictors via random

slopes ("v"), and to cluster-specific means via intercept variance ("m"). Given that no variation is explained by "f" or "v" in this null model, in this specific situation "fvm" is equal to "m" in the R-squared output. The remaining 68.4% of variance is residual variance, shown as "sigma2" in the decomposition output and "residual (within)" in the total bar graph.

We can double-check the results by manually calculating the ICC, which describes the proportion of variability in the outcome accounted for by cluster membership, and is equivalent conceptually and mathematically to $R^{2(m)}_{total}$ in the special case of the random-intercept-only model. The ICC is calculated as $ICC = \frac{between\ variance}{between\ variance + within\ variance}$. Given the model output generated above with the call *summary(null_model)*, we calculate the ICC as follows:

```
0.699 / (0.699 + 1.516)

## [1] 0.3155756
```

With an ICC of 0.316, 31.6% of the variation in teacher job satisfaction can be attributed to school membership, matching the output of *r2mlm*.

Level 1 fixed effects

As we just saw, including a random intercept can account for total and between variance, but no within variance. To explain within variance, we need to include level 1 predictors. To demonstrate, we'll now include fixed effects for the level 1 predictors of school-mean-centered teacher salary (*salary_c*) and school-mean-centered perceived control over the curriculum (*control_c*). This model assesses whether teacher salary and control over curriculum are related to job satisfaction within school. We'll consider the fixed effects now, then add random slopes in the next model.

Level 1: $satisfaction_{ij} = \beta_{0j} + \beta_{1j} * salary\_c_{ij} + \beta_{2j} * control\_c_{ij} + R_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + U_{0j}$

$\beta_{1j} = \gamma_{10}$

$\beta_{2j} = \gamma_{20}$

Combined: $satisfaction_{ij} = \gamma_{00} + \gamma_{10} * salary\_c_{ij} + \gamma_{20} * control\_c_{ij} + U_{0j} + R_{ij}$

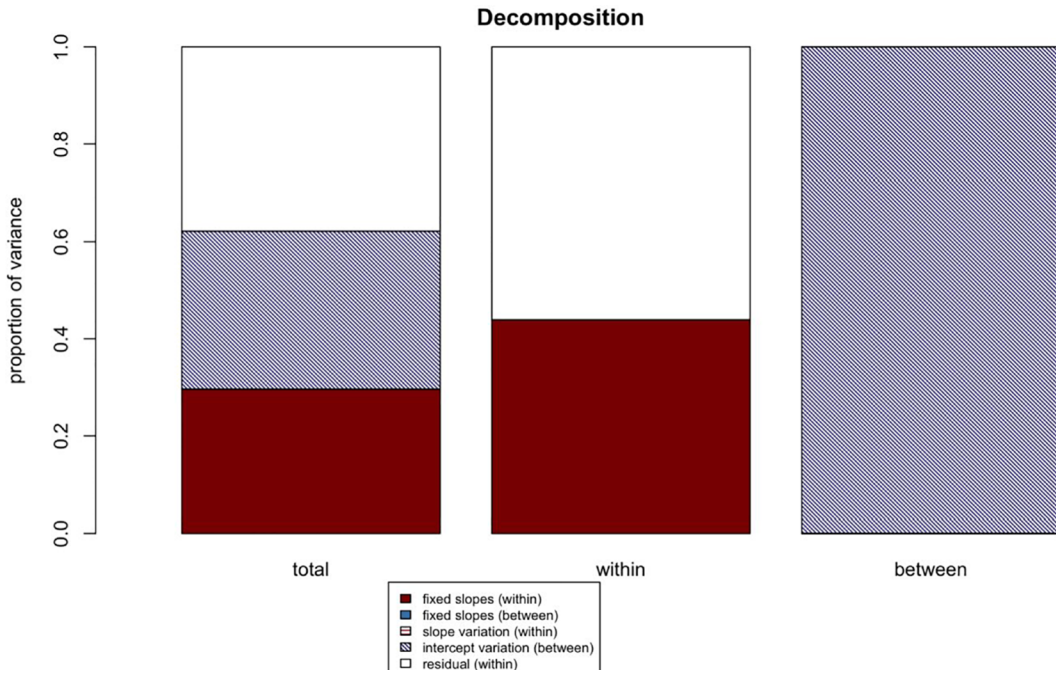
```
l1_model_fixed <- lmer(satisfaction ~ 1 + salary_c + control_c + (1 | schoolID),
                       data = teachsat,
                       REML = TRUE)
summary(l1_model_fixed)

## Linear mixed model fit by REML ['lmerMod']
## Formula: satisfaction ~ 1 + salary_c + control_c + (1 | schoolID)
##    Data: teachsat
##
## REML criterion at convergence: 24962.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.6362 -0.6375  0.0057  0.6464  3.6596
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  schoolID (Intercept) 0.7215   0.8494
##  Residual             0.8384   0.9156
## Number of obs: 9000, groups:  schoolID, 300
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 5.996774   0.049983  119.98
## salary_c    0.074007   0.001115   66.40
## control_c   0.310644   0.006104   50.89
##
## Correlation of Fixed Effects:
##           (Intr) slry_c
## salary_c   0.000
## control_c  0.000 -0.005
```

Per the model summary of fixed effects, the estimated intercept for job satisfaction is 6.00 on a 1 to 10 scale; because both predictors have a mean of 0, we can interpret this intercept as the estimated grand mean of satisfaction, as well as the predicted value of satisfaction at the mean of the predictors. For a one-unit (i.e., thousand-dollar) increase in salary relative to the school mean, predicted satisfaction increases by 0.07 units, holding curriculum control constant. For a one-unit increase in curriculum control relative to the school mean, predicted satisfaction increases by 0.31 units, holding salary constant. Per the model summary of random effects, the predicted between-school intercept variance is 0.72. The estimated within-school residual variation resulting from individual variation of teachers around their school's predicted mean job satisfaction is 0.84.

To calculate effect sizes for this model, we run:

```
r2mlm(l1_model_fixed)
```




```
## $Decompositions
##                  total              within             between
## fixed, within    0.295843451292118 0.438746423745784 NA
## fixed, between   0                 NA                0
## slope variation  0                 0                 NA
## mean variation   0.325707435364683 NA                1
## sigma2           0.378449113343199 0.561253576254216 NA
##
## $R2s
##      total              within             between
## f1   0.295843451292118 0.438746423745784 NA
## f2   0                 NA                0
## v    0                 0                 NA
## m    0.325707435364683 NA                1
## f    0.295843451292118 NA                NA
## fv   0.295843451292118 0.438746423745784 NA
## fvm  0.621550886656801 NA                NA
```

For the null model, the only component accounting for variance in job satisfaction was intercept variation. With the addition of level 1 predictors, we can consider total and/or within-cluster variance explained by level 1 predictors via their fixed effects. This is denoted "fixed, within" in the decomposition output, "f1" in the R-squared output, and "fixed slopes (within)" in the graphical output. The level 1 predictors explain an estimated 29.6% of the total variance (the total column of decompositions and R-squareds) and 43.9% of within variance (the within column) via their fixed slopes. We can also see that the level 1 predictor via fixed slopes ("f") and the cluster means via intercept variance ("m") in combination account for 62.2% of total variance with the "fvm" term. Recall that we haven't yet added random slope variation ("v"). In the "fvm" term, no variance is presently explained by "v" because no level 1 predictor yet has contributed to explained variance via random slope variation. Between variance is unaffected by the addition of the level 1 predictors, because they vary exclusively within-cluster and hence cannot explain between-cluster variation.

The *r2mlm()* output describes variance explained by all level 1 predictors via the fixed effects. If we wanted to examine the *unique* contributions of each individual fixed effect, we would compare models using the *r2mlm_comp()* function. We demonstrate this functionality later.

Level 1 fixed and random effects

Suppose our theory suggests that the effect of curriculum control on job satisfaction varies across schools. To allow for such variation, we can add a random slope for curriculum control to the model, represented by U_{2j} in the equation for β_{2j}.

$$Level\ 1: satisfaction_{ij} = \beta_{0j} + \beta_{1j} * salary\_c_{ij} + \beta_{2j} * control\_c_{ij} + R_{ij}$$
$$Level\ 2: \beta_{0j} = \gamma_{00} + U_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20} + U_{2j}$$

Combined: $satisfaction_{ij} = \gamma_{00} + \gamma_{10} * salary\_c_{ij} + \gamma_{20} * control\_c_{ij} + U_{0j} + U_{2j} * control\_c_{ij} + R_{ij}$

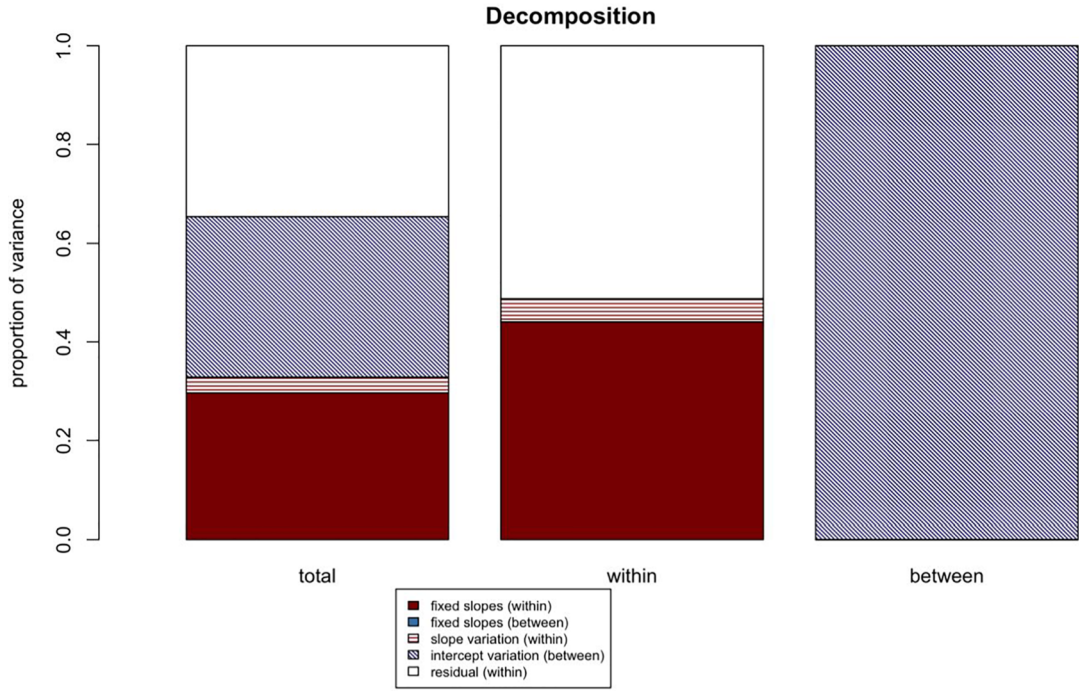
```
l1_model_random <- lmer(satisfaction ~ 1 + salary_c + control_c + (1 + control_c | schoolID),
                        data = teachsat,
                        REML = TRUE)
summary(l1_model_random)
## Linear mixed model fit by REML ['lmerMod']
## Formula: satisfaction ~ 1 + salary_c + control_c + (1 + control_c | schoolID)
##    Data: teachsat
##
## REML criterion at convergence: 24565.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.5976 -0.6300  0.0074  0.6395  3.7882
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  schoolID (Intercept) 0.72400  0.8509
##           control_c   0.02826  0.1681   0.07
##  Residual             0.76561  0.8750
## Number of obs: 9000, groups:  schoolID, 300
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 5.996774   0.049984  119.97
## salary_c    0.074135   0.001078   68.75
## control_c   0.311281   0.011361   27.40
##
## Correlation of Fixed Effects:
##           (Intr) slry_c
## salary_c   0.000
## control_c  0.058 -0.004
```

The fixed effects have the same interpretation as in the last model, with the exception that the slope of *control_c* now represents the across-cluster *average* slope. In this model, we newly introduced a random effect for *control_c*: the estimated across-school variance in the slope of curriculum control is 0.03. The estimated across-school intercept variance is 0.72 and the estimated within-school residual variance is 0.77.

With *r2mlm* we can consider the impact of adding a random effect of curriculum control on variance explained.

The impact of the level 1 predictor via its random slope is denoted "slope variation" in the decompositions output, "v" in the R-squared output, and "slope variation (within)" in the graphical output. This added random slope accounts for 3.2% of total variance and 4.7% of within variance. The between variance explained is unaffected by the addition of the random slope, as the level 1 variable curriculum control varies exclusively within cluster and hence cannot explain between-cluster variance.

```
r2mlm(l1_model_random)
```




```
## $Decompositions
##                 total             within             between
## fixed, within   0.296567159372403 0.440275210302958  NA
## fixed, between  0                 NA                 0
## slope variation 0.031859752893799 0.0472980873377948 NA
## mean variation  0.326405047496696 NA                 1
## sigma2          0.345168040237102 0.512426702359247  NA
##
## $R2s
##     total             within             between
## f1  0.296567159372403 0.440275210302958  NA
## f2  0                 NA                 0
## v   0.031859752893799 0.0472980873377948 NA
## m   0.326405047496696 NA                 1
## f   0.296567159372403 NA                 NA
## fv  0.328426912266202 0.487573297640753  NA
## fvm 0.654831959762898 NA                 NA
```

Level 2 fixed effects

By adding level 1 effects to our model, we have been considering factors that relate to job satisfaction within schools. For example, "Within a school, how are salary and curriculum control related to job satisfaction?" and "To what extent does curriculum control relate to job satisfaction differently across schools?" Now, by adding level 2 predictors to the model, we can assess how school-level factors may affect job satisfaction. For our example, we'll add student–teacher ratio, with higher values indicating more students per teacher. This variable does not vary within schools, only between schools, and hence will only explain between-school variance. That is, each school has only one value for student–teacher ratio.

Level 1: $satisfaction_{ij} = \beta_{0j} + \beta_{1j} * salary\_c_{ij} + \beta_{2j} * control\_c_{ij} + R_{ij}$
Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01} * s\_t\_ratio_j + U_{0j}$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20} + U_{2j}$$

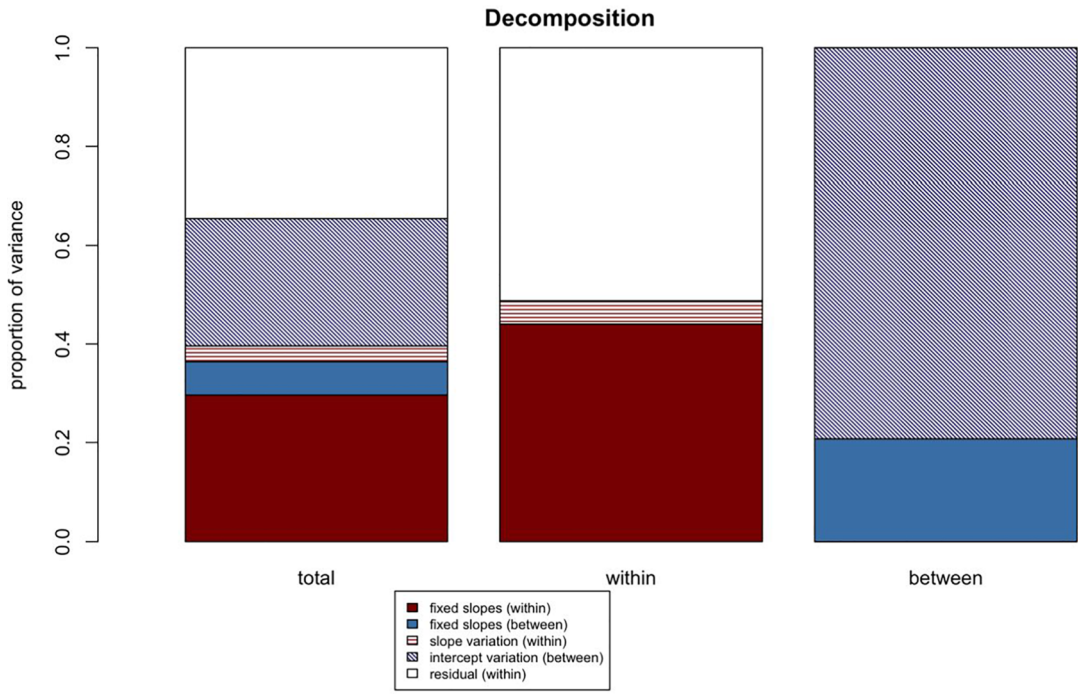
Combined: $satisfaction_{ij} = \gamma_{00} + \gamma_{01} * s\_t\_ratio_j + \gamma_{10} * salary\_c_{ij} + \gamma_{20} * control\_c_{ij} + U_{0j} + U_{2j} * control\_c_{ij} + R_{ij}$

```
l2_model <- lmer(satisfaction ~ 1 + control_c + salary_c + s_t_ratio + (1 + control_c |
schoolID),
                 data = teachsat,
                 REML = TRUE)
summary(l2_model)

## Linear mixed model fit by REML ['lmerMod']
## Formula:
## satisfaction ~ 1 + control_c + salary_c + s_t_ratio + (1 + control_c |
##     schoolID)
##    Data: teachsat
##
## REML criterion at convergence: 24507.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.6115 -0.6275  0.0108  0.6414  3.7958
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  schoolID (Intercept) 0.57478  0.7581
##           control_c   0.02826  0.1681   0.07
##  Residual             0.76561  0.8750
## Number of obs: 9000, groups:  schoolID, 300
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  7.186462   0.144236  49.824
## control_c    0.311279   0.011361  27.398
## salary_c     0.074132   0.001078  68.752
## s_t_ratio   -0.037178   0.004285  -8.676
##
## Correlation of Fixed Effects:
##           (Intr) cntrl_ slry_c
## control_c  0.017
## salary_c   0.000 -0.004
## s_t_ratio -0.951  0.000  0.000
```

For an increase of one student per teacher, there is a 0.04-unit decrease in predicted teacher job satisfaction, controlling for the other effects in the model. With *r2mlm*, we can consider the impact of adding this level 2 predictor on variance explained.

```
r2mlm(l2_model)
```




```
## $Decompositions
##                   total              within             between
## fixed, within    0.296431806719555  0.440263091958242  NA
## fixed, between   0.0676695868874132 NA                 0.207134501406624
## slope variation 0.0318477856338068 0.0473006076180897 NA
## mean variation   0.259024355588986  NA                 0.792865498593376
## sigma2           0.34502646517024   0.512436300423669  NA
##
## $R2s
##     total              within             between
## f1  0.296431806719555  0.440263091958242  NA
## f2  0.0676695868874132 NA                 0.207134501406624
## v   0.0318477856338068 0.0473006076180897 NA
## m   0.259024355588986  NA                 0.792865498593376
## f   0.364101393606968  NA                 NA
## fv  0.395949179240775  0.487563699576332  NA
## fvm 0.65497353482976   NA                 NA
```

The impact of the level 2 predictor via its fixed effect is denoted "fixed, between" in the decompositions output, "f2" in the R-squared output, and "fixed slopes (between)" in the graphical output. Student–teacher ratio explains 6.8% of total variance and 20.7% of between-school variance in teacher job satisfaction via its fixed effect. The level 1 and level 2 predictors together now explain 36.4% of total variance via fixed effects, captured by the "f" term of the R-squared output.

Model comparisons

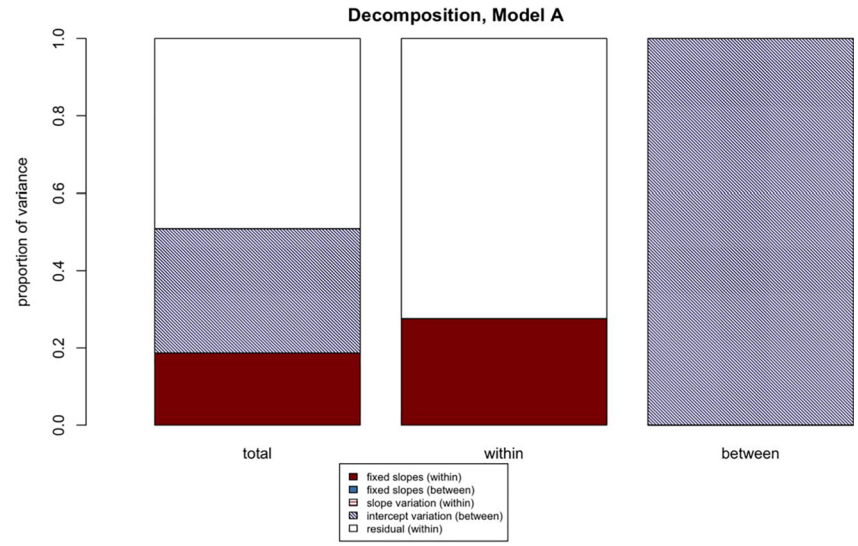
Earlier, we added fixed effects for two level 1 predictors – salary and curriculum control – to our model at the same time. We noted that doing so does not tell us how much variance each effect explains *uniquely*. One way we can assess unique contributions of individual predictors to variance explained is by comparing two models: one model without the predictor of interest and one model with the predictor of interest. The first step to getting the associated effect sizes is to run these models using *lmer* or *nlme*.

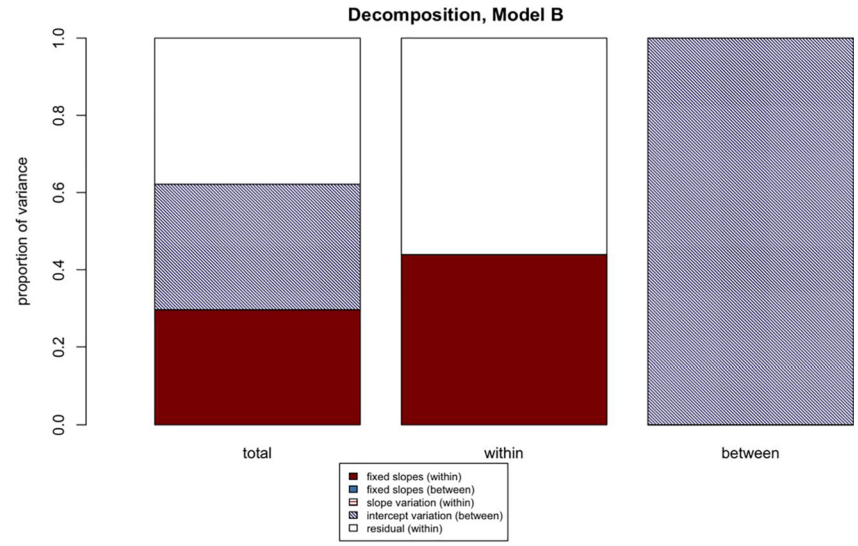
The single-effect model *model_salary* will yield variance explained by salary alone, the model with both effects *model_both* will yield variance explained by both effects (which we calculated earlier), and the difference between the two models in $R^{2(f1)}_{total}$ and $R^{2(f1)}_{within}$ will yield, respectively, the total and the within variance uniquely explained by curriculum control over and above salary. We can compare the models using the *r2mlm_comp()* function, which takes two models as arguments. Because we are interested in assessing the contributions of the predictors via their fixed effects, we will focus on the difference in R-squared measures that have "f1" as their source of explained variance (see Rights & Sterba, 2020). The graphical output for this function includes five plots: (1) decomposition of between-cluster variance for both Model A and Model B; (2) decomposition of within-cluster variance for both Model A and Model B; (3) decomposition of total variance for both Model A and Model B; (4) full decomposition for Model A; and (5) full decomposition for Model B. Note that for brevity we only explain (4) and (5), the overall decomposition plots.

```
# Single-effect model, just salary_c
model_salary <- lmer(satisfaction ~ 1 + salary_c + (1 | schoolID),
                     data = teachsat,
                     REML = TRUE)

# Model with both effects (the same as l1_model_fixed from earlier)
model_both <- lmer(satisfaction ~ 1 + salary_c + control_c + (1 | schoolID),
                     data = teachsat,
                     REML = TRUE)
```

```
r2mlm_comp(model_salary, model_both)
```






```
## $`Model A R2s`
##     total             within            between
## f1  0.186927533490092 0.275690579940632 NA
## f2  0                 NA                0
## v   0                 0                 NA
## m   0.321966192931418 NA                1
## f   0.186927533490092 NA                NA
## fv  0.186927533490092 0.275690579940632 NA
## fvm 0.50889372642151  NA                NA
##
## $`Model B R2s`
##     total             within            between
## f1  0.295843451292118 0.438746423745784 NA
## f2  0                 NA                0
## v   0                 0                 NA
## m   0.325707435364683 NA                1
## f   0.295843451292118 NA                NA
## fv  0.295843451292118 0.438746423745784 NA
## fvm 0.621550886656801 NA                NA
##
## $`R2 differences, Model B - Model A`
##           total    within between
## f1  0.108915918 0.1630558      NA
## f2  0.000000000         NA       0
## v   0.000000000 0.0000000      NA
## m   0.003741242         NA       0
## f   0.108915918         NA      NA
## fv  0.108915918 0.1630558      NA
## fvm 0.112657160         NA      NA
```

In our case, Model A is *model_salary*, so the "Model A R2s" output describes the variance explained by salary by itself. Roughly 18.7% of total variance and 27.6% of within-school variance in teacher job satisfaction is explained by teacher salary via its fixed effect. Model B is *model_both*, so the "Model B R2s" output describes variance explained by both salary and curriculum control; this matches the earlier *l1_model_fixed* output: both level 1 predictors explain 29.6% of total variance and 43.9% of within-school variance in job satisfaction via the fixed effects. The variance uniquely explained by curriculum control accounts for the difference between the one-effect model and the both-effects model, and is described in the "R2 differences, Model B - Model A" output. Curriculum control uniquely explains 10.9% of total variance and 16.3% of the within-school variance in job satisfaction via its fixed effect.

Note that if the models being compared are not nested, you also need to provide your data: *r2mlm_comp(modelA, modelB, data)*. For more on comparing models, including an elaboration on different strategies and the appropriate R-squared difference measure to use for each possible type of model comparison, see Rights and Sterba (2020).

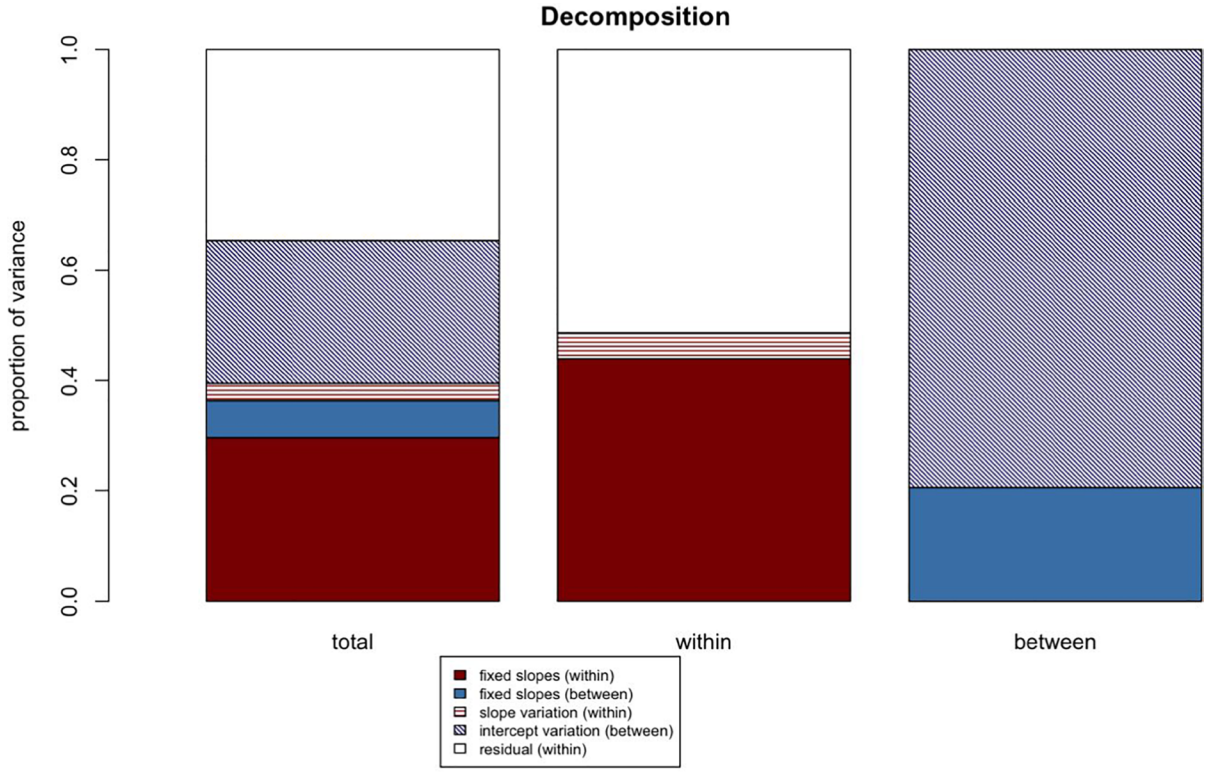
Manual entry

If you used another software to run MLMs (e.g., MPlus, SPSS) and not *lme4* or *nlme* in R, then you can manually enter information about your model and dataset to calculate R-squared estimates using *r2mlm_manual*, which takes the following parameters as input:

- *data*: your dataset
- *within_covs*: list of numbers or variable names corresponding to the column numbers or variable names in your dataset for level 1 predictors
- *between_covs*: list of numbers or variable names corresponding to the column numbers or variable names in your dataset for level 2 predictors
- *random_covs*: list of numbers or variable names corresponding to the column numbers or variable names in your dataset for level 1 predictors with random effects
- *gamma_w*: list of fixed slope estimates for level 1 predictors in the order listed in *within_covs*
- *gamma_b*: list of intercept estimate (if applicable) followed by fixed slope estimates for level 2 predictors in the order listed in *between_covs*
- *Tau*: random effect covariance matrix. The first row/column denotes the intercept variances and covariances; set to 0 if intercept is fixed. Subsequent rows/columns denote random slope variances and covariances in the order listed in *random_covs*
- *sigma2*: level 1 residual variance
- *has_intercept*: true/false indicating whether your model estimates an intercept; default value of true
- *clustermeancentered*: true/false indicating whether your level 1 predictors are centered-within-cluster; default value of true

Manual entry for *l2_model* would look as follows:

```
r2mlm_manual(data = teachsat,
             within_covs = c(4, 5),
             between_covs = c(8),
             random_covs = c(4),
             gamma_w = c(0.311, 0.074),
             gamma_b = c(7.186, -0.037),
             Tau = matrix(c(0.575, 0.009, 0.009, 0.028), 2, 2),
             sigma2 = 0.766,
             has_intercept = TRUE,
             clustermeancentered = TRUE)
```




```
## $Decompositions
##                  total              within             between
## fixed, within    0.296022422439019  0.439625147440015  NA
## fixed, between   0.0671264807082975 NA                 0.205500898247648
## slope variation 0.0316015152901635 0.0469316503266841 NA
## mean variation   0.259521632661036  NA                 0.794499101752352
## sigma2           0.345727948901484  0.513443202233301  NA
##
## $R2s
##      total              within             between
## f1   0.296022422439019  0.439625147440015  NA
## f2   0.0671264807082975 NA                 0.205500898247648
## v    0.0316015152901635 0.0469316503266841 NA
## m    0.259521632661036  NA                 0.794499101752352
## f    0.363148903147317  NA                 NA
## fv   0.39475041843748   0.486556797766699  NA
## fvm 0.654272051098516  NA                 NA
```

Excepting some trivial differences due to rounding the input values, these results match those calculated with *r2mlm(l2_model)*. A similar manual entry process is possible for comparing models using *r2mlm_comp_manual()*.

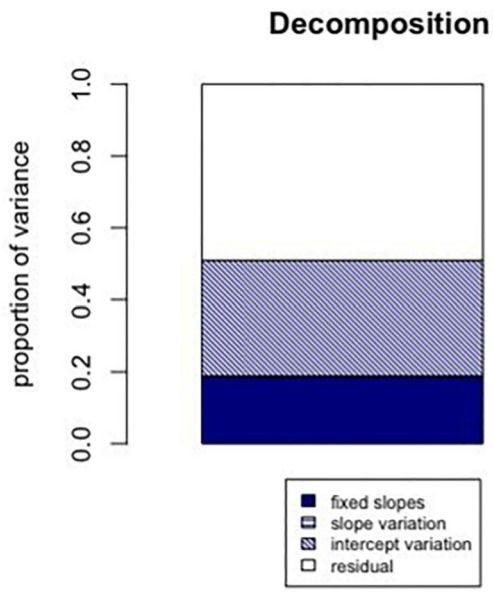
Models with non-cluster-mean-centered level 1 predictors

Researchers do not always wish to cluster-mean-center level 1 predictors. For example, in longitudinal contexts in which "time" is a level 1 predictor, researchers might want to center "time" at the first measurement occasion rather than at a person's mean time. If a researcher's level 1 predictors are not all cluster-mean-centered, the *r2mlm* package provides two options for calculating R-squared values: the *r2mlm()* function and the *r2mlm_long_manual()* function. To demonstrate both options, we will first remove the cluster-mean-centering from *salary_c* by adding a constant to every value. We will then run a model predicting *satisfaction* by *salary* (uncentered).

```
teachsat$salary <- teachsat$salary_c + 2
uncentered_model <- lmer(satisfaction ~ salary + (1 | schoolID), data = teachsat)
```

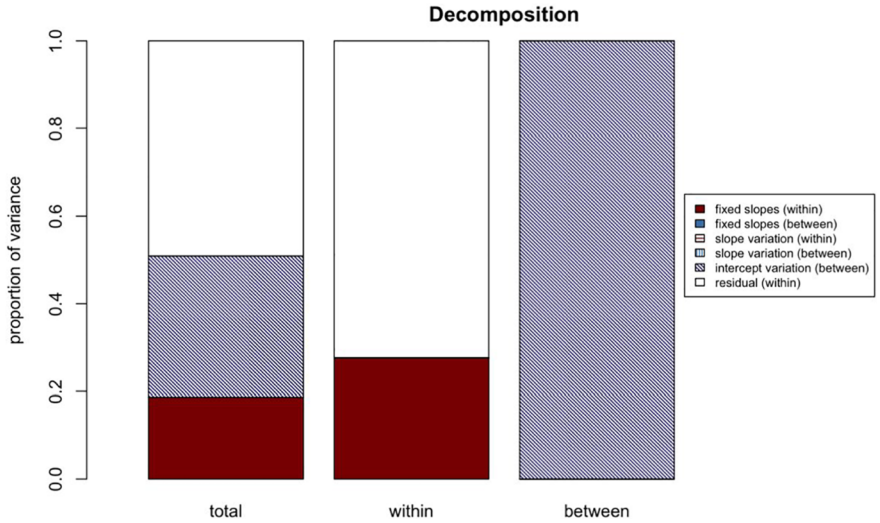
The *r2mlm()* function calculates a decomposition of variance yielding total measures.

```
r2mlm(uncentered_model)
```




```
## $Decompositions
##                     total
## fixed           0.1869275
## slope variation 0.0000000
## mean variation  0.3219662
## sigma2          0.4911063
##
## $R2s
##         total
## f   0.1869275
## v   0.0000000
## m   0.3219662
## fv  0.1869275
## fvm 0.5088937
```

Alternatively, the *r2mlm_long_manual()* function calculates both a total decomposition of variance and level-specific decompositions of variance, yielding total, within-cluster, and between-cluster measures.




```
## $Decompositions
##                                 total
## fixed slopes (within)         0.186907226965456
## fixed slopes (between)        0.000000000000000000000000000000000761887047387227
## slope variation (within)      0
## slope variation (between)     0
## intercept variation (between) 0.321974234044723
## residual (within)             0.491118538989821
##                                 within
## fixed slopes (within)         0.275663900032059
## fixed slopes (between)        NA
## slope variation (within)      0
## slope variation (between)     NA
## intercept variation (between) NA
## residual (within)             0.724336099967941
##                                 between
## fixed slopes (within)         NA
## fixed slopes (between)        0.000000000000000000000000000000000236629819043657
## slope variation (within)      NA
## slope variation (between)     0
## intercept variation (between) 1
## residual (within)             NA
##
## $R2s
##     total                                                 within
## f1  0.186907226965456                                     0.275663900032059
## f2  0.000000000000000000000000000000000761887047387227 NA
## v1  0                                                     0
## v2  0                                                     NA
## m   0.321974234044723                                     NA
## f   0.186907226965456                                     NA
## fv  0.186907226965456                                     0.275663900032059
## fvm 0.508881461010179                                     NA
##     between
## f1  NA
## f2  0.000000000000000000000000000000000236629819043657
## v1  NA
## v2  0
## m   1
## f   NA
## fv  0.000000000000000000000000000000000236629819043657
## fvm NA
```

See Rights and Sterba (2021) for a demonstration and more information about using *r2mlm_long_manual()* to calculate R-squareds for models with heteroscedastic variance estimates. Note that an automatic *r2mlm_long()* function is under active development.

Discussion

Reporting effect sizes is necessary to contextualize results. Rights and Sterba (2019) developed a comprehensive effect size framework for R-squared in MLMs that integrates previously developed MLM R-squareds as special cases, and Shaw et al. (2020) implemented the framework into an accessible R package, *r2mlm*. In this tutorial, we demonstrated how to use and interpret output from *r2mlm*. We will now discuss considerations for appropriately reporting results, package limitations, and future directions.

Appropriate reporting

The most important consideration when reporting and interpreting R-squared values is context: they should be reported in the context of other model information, and understood in the context of the data at hand, how the variables were measured, and the relevant literature. As a standardized effect size, R-squared has advantages and drawbacks (Pek & Flora, 2018). Advantageously, it has an intuitive zero-to-one range regardless of the measures involved. This standardization facilitates interpreting results for measures that do not have meaningful units. However, standardized metrics are calculated based on the variability of the sample. As such, they cannot necessarily be compared across samples that have substantially different degrees of variation in the outcome and/or the predictors. Additionally, some unstandardized metrics do have interpretable units that provide valuable insights related to a research question. As exemplified in the above data demonstration, one should interpret raw MLM parameter estimates alongside standardized R-squared effect sizes – as well as additional information like significance of and precision of the estimates – to yield a full picture of one's results.

Assessing the size of an R-squared value is also a context-specific exercise. The cutoffs for R-squared values proposed in Cohen (1988) are sometimes treated as global recommendations. However, Cohen noted that his cutoffs were suggestions that should be rejected if they are "unsuited to the substantive content of any given investigation" (p. 414). The takeaway from his recommendations was that small, medium, and large benchmarks were for a given context, and researchers should consider their R-squared measures in the context of the relevant literature and their theory. The interpretation of a given R-squared should be tempered by considerations like sample size, measures involved, and the nature of a manipulation (Cortina & Landis, 2009).

Finally, we will note that effect sizes are part of a toolbox of rigorous research practices that also includes transparent reporting and valid measurement. To complement the greater flexibility afforded by our *r2mlm* R package regarding what R-squared measures to report, we recommend that researchers also preregister their study and include mention of the effect sizes they will report and what sizes they expect or consider large given the context.

Relation to other R packages

There are other R packages and functions dedicated to estimating multilevel R-squared values, but none that provide the full partitioning of variance that *r2mlm* does. The *multilevelR2* function from the *mitml* package (Grund et al., 2021) allows users to calculate the MLM R-squared values proposed by Raudenbush and Bryk (2002), Snijders and Bosker (2011), and LaHuis et al. (2014). The *r2glmm* package (Jaeger, 2017) implements R-squared measures described in Johnson (2014), Jaeger et al. (2017), and Edwards et al. (2008). In the linear mixed model framework, the functionality of both packages is subsumed by *r2mlm*. *r2glmm* does extend the three sets of measures from Johnson (2014), Jaeger et al. (2017), and Edwards et al. (2008) to a generalized linear mixed model framework with, for instance, binary outcomes. For mapping these special cases onto notation from the general framework of measures, see Table 3 in Rights and Sterba (2019); for discussion on the relation between the general framework and the Jaeger et al. (2017) and Edwards et al. (2008) measures, see Rights and Sterba (2019, 2020).

Future directions and limitations

Effect sizes for MLMs are the subject of active methodological research. As evidenced by the breadth of the Rights and Sterba (2019) framework and the number of independently developed R-squared measures predating and subsumed by it, there are a number of different ways of decomposing explained variance for a multilevel model. One can break total variance down by more general categories of source contribution (i.e., contributions of all predictors via fixed effects – source f1+f2 – vs. contribution of all random effects – source m+v) or further by individual source type (e.g., contributions of predictors via level 1 fixed effects – f1 – versus contributions of predictors via level 2 fixed effects – f2). Furthermore, there are multiple ways of quantifying the contribution of individual predictors. Rights and Sterba

(2020) discuss a simultaneous approach in which R-squared differences between models quantify proportions of variance explained by individual terms over and above *all other* terms, as well as a hierarchical approach in which R-squared differences quantify the proportion of variance explained by individual terms over and above the *previously entered* terms. An alternative approach, known as Shapley regression (Shapley, 1953) or dominance analysis (Budescu, 1993; Azen & Budescu, 2003), involves considering all possible subset models to which a predictor could be added and computing the average change in an R-squared measure arising from adding the target predictor (versus another predictor) into each different possible subset model. This approach has been extended from the single-level to the multilevel context using a subset of possible R-squared difference measures (Luo & Azen, 2013), and in future work can be further extended to use the full suite of R-squared difference measures from Rights and Sterba (2020).

As illustrated, the *r2mlm* R package can be used to calculate effect sizes for two-level MLMs. Functionality for three-or-more-level models (Rights & Sterba, in press) is in development. There is currently a manual entry option for three-level models; see help(r2mlm3) for documentation. Functionality for other model complexities is also in development. There is currently a manual entry option for models with heteroscedastic and/or autocorrelated level 1 residuals, which also provides level-specific variance explained under any centering option; see help(r2mlm_long_manual) for documentation, and Rights and Sterba (2021) for more details. Cross-classified models are not currently supported. You cannot use the I() function within a model to create higher-order terms; such terms need to exist explicitly as variables in your dataset. Additionally, only your cluster variable can be categorical in the dataset; all other variables in the model must be numeric (thus, to incorporate categorical predictors, one must directly input the associated independent variable codes, e.g., dummy or effects codes). The *r2mlm* package is under active development. If any interested readers would like to report a bug or make a request for functionality, they can file an issue or pull request at www.github.com/mkshaw/r2mlm.

Availability of Data The data analyzed in the above manuscript are available as part of the *r2mlm* package, https://CRAN.R-project.org/package=r2mlm.

Code Availability All software and code used in the above manuscript is provided in the manuscript itself.

Authors' Contributions MS and JKF developed the manuscript structure. MS and JDR created the illustrative dataset and wrote and tested the R functions. MS coded the examples. All authors contributed to writing the manuscript and the R package documentation.

Funding This research was supported by an Association for Psychological Science small grant, the Fonds de recherché du Québec – Nature et technologies, a Natural Sciences and Engineering Research Council of Canada Discovery Grant and Discovery Launch Supplement.

Declarations

Conflict of Interest/Competing Interests The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

Aguinis, H., & Culpepper, S. A. (2015). An expanded decision-making procedure for examining cross-level interaction effects with multilevel modeling. *Organizational Research Methods, 18*, 155–176. https://doi.org/10.1177/1094428114563618

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods, 8*(2), 129–148. https://doi.org/10.1037/1082-989X.8.2.129

Baruffa, O. (2021). *Big Book of R.* http://bigbookofr.com

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bickel, R. (2007). *Multilevel analysis for applied research. It's just regression!* New York, NY: Guilford Press.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin, 114*(3), 542–551. https://doi.org/10.1037/0033-2909.114.3.542

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciencesa* (2nd ed.): Lawrence Erlbaum Associates.

Cortina, J. M., & Landis, R. S. (2009). When small effect sizes tell a big story, and when large effect sizes don't *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences.* (pp. 287-308). New York, NY, US: Routledge/Taylor & Francis Group.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Demidenko, E., Sargent, J., & Onega, T. (2012). Random effects coefficient of determination for mixed and meta-analysis models. *Communications in Statistics Theory and Methods, 41*, 953–969. https://doi.org/10.1080/03610926.2010.535631

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., & Schabenberger, O. (2008). An R^2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine, 27*, 6137–6157. https://doi.org/10.1002/sim.3429

Enders, C., & Tofighi, D. (2007). Centering Predictor Variables in Cross-Sectional Multilevel Models: A New Look at An Old Issue. *Psychological Methods, 12*, 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Finch, W. H., Bolin, J. E., & Kelley, K. (2014). *Multilevel Modeling Using R*: Taylor & Francis.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical Methods in Education and Psychology*: Allyn and Bacon.

Grund, S., Robitzsch, A., & Luedtke, O. (2021). mitml: Tools for Multiple Imputation in Multilevel Modeling. R package version 0.4-3. https://CRAN.R-project.org/package=mitml

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications, 2nd ed*. New York, NY, US: Routledge/Taylor & Francis Group.

Jaeger, B. C. (2017). r2glmm: Computes R Squared for Mixed (Multilevel) Models. R package version 0.1.2. https://CRAN.R-project.org/package=r2glmm

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics, 44*, 1086–1105. https://doi.org/10.1080/02664763.2016.1193725

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution, 5*(9), 944–946. https://doi.org/10.1111/2041-210X.12225

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*(2), 137–152. https://doi.org/10.1037/a0028086

Kramer, M. (2005). R2 statistics for mixed models. *2005 Proceedings of the Conference on Applied Statistics in Agriculture* (pp. 148–160). Manhattan, KS: Kansas State University.

Kreft, I. G., & de Leeuw, J. (1998). *Introducing Multilevel Modeling*. https://doi.org/10.4135/9781849209366

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained Variance Measures for Multilevel Models. *Organizational Research Methods, 17*(4), 433–451. https://doi.org/10.1177/1094428114541701

Luo, W., & Azen, R. (2013). Determining Predictor Importance in Hierarchical Linear Models Using Dominance Analysis. *Journal of Educational and Behavioral Statistics, 38*(1), 3–31. https://doi.org/10.3102/1076998612458319

McCoach, D. B., & Adelson, J. L. (2010). Dealing With Dependence (Part I): Understanding the Effects of Clustered Data. *Gifted Child Quarterly, 54*(2), 152–155. https://doi.org/10.1177/0016986210363076

McCoach, D. B. (2010). Dealing With Dependence (Part II): A Gentle Introduction to Hierarchical Linear Modeling. *Gifted Child Quarterly, 54*(3), 252–256. https://doi.org/10.1177/0016986210373475

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Orelien, J. G., & Edwards, L. J. (2008). Fixed-effect variable selection in linear mixed models using R2 statistics. *Computational Statistics & Data Analysis, 52*(4), 1896–1907.

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods, 23*(2), 208–225. https://doi.org/10.1037/met0000126

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team (2020). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1–144, https://CRAN.R-project.org/package=nlme

Psychonomic Society (2012). *Statistical Guidelines*. Psychonomic Society. https://www.psychonomic.org/page/statisticalguidelines

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*: SAGE Publications.

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods, 24*(3), 309–338. https://doi.org/10.1037/met0000184

Rights, J. D., & Sterba, S. K. (2020). New Recommendations on the Use of R-Squared Differences in Multilevel Model Comparisons. *Multivariate Behavioral Research, 55*(4), 568–599. https://doi.org/10.1080/00273171.2019.1660605

Rights, J. D., & Sterba, S. K. (2021). Effect size measures for longitudinal growth analyses: Extending a framework of multilevel model R-squareds to accommodate heteroscedasticity, autocorrelation, nonlinearity, and alternative centering strategies. *New directions for child and adolescent development, 2021*(175), 65–110. https://doi.org/10.1002/cad.20387

Rights, J. D., & Sterba, S. K. (in press). R-squared measures for multilevel models with three or more levels. *Multivariate Behavioral Research.*

Shapley, L. S. (1953). Contributions to the Theory of Games (AM-28), Volume II. In K. Harold William & T. Albert William (Eds.), *17. A Value for n-Person Games* (pp. 307–318). Princeton University Press.

Shaw, M., Rights, J. D., Sterba, S. K., & Flake, J. K. (2020). *r2mlm: R-Squared Measures for Multilevel Models*. https://doi.org/10.31234/osf.io/xc4sv

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled Variance in Two-Level Models. *Sociological Methods & Research, 22*(3), 342–363. https://doi.org/10.1177/0049124194022003004

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* London: SAGE Publications.

Snijders, T. A. B., & Bosker, R. J. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. London, UK: Sage.

Vonesh, E. F., & Chinchilli, V. M. (1997). *Linear and nonlinear models for the analysis of repeated measurements*. New York, NY: Marcel Dekker.

Wang, J., & Schaalje, G. B. (2009). Model Selection for Linear Mixed Models Using Predictive Criteria. *Communications in Statistics - Simulation and Computation, 38*(4), 788–801. https://doi.org/10.1080/03610910802645362

Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*: O'Reilly Media.

Wright, S. (1921). Correlation and Causation. *Journal of Agricultural Research, 20*, 557–585.

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in Medicine, 22*, 3527–3541. https://doi.org/10.1002/sim.1572

Open Practices Statement

The data and materials analyzed in the manuscript are available as part of the *r2mlm* package, https://CRAN.R-project.org/package=r2mlm.