Routledge
Taylor & Francis Group

Check for updates

New Recommendations on the Use of R-Squared Differences in Multilevel Model Comparisons

Jason D. Rights[a] and Sonya K. Sterba[b]

[a]University of British Columbia; [b]Vanderbilt University

ABSTRACT

When comparing multilevel models (MLMs) differing in fixed and/or random effects, researchers have had continuing interest in using R-squared differences to communicate effect size and importance of included terms. However, there has been longstanding confusion regarding which R-squared difference measures should be used for which kind of MLM comparisons. Furthermore, several limitations of recent studies on R-squared differences in MLM have led to misleading or incomplete recommendations for practice. These limitations include computing measures that are by definition incapable of detecting a particular type of added term, considering only a subset of the broader class of available R-squared difference measures, and incorrectly defining what a given R-squared difference measure quantifies. The purpose of this paper is to elucidate and resolve these issues. To do so, we define a more general set of total, within-cluster, and between-cluster R-squared difference measures than previously considered in MLM comparisons and give researchers concrete step-by-step procedures for identifying which measure is relevant to which model comparison. We supply simulated and analytic demonstrations of limitations of previous MLM studies on R-squared differences and show how application of our step-by-step procedures and general set of measures overcomes each. Additionally, we provide and illustrate graphical tools and software allowing researchers to automatically compute and visualize our set of measures in an integrated manner. We conclude with recommendations, as well as extensions involving (a) how our framework relates to and can be used to obtain pseudo-R-squareds, and (b) how our framework can accommodate both simultaneous and hierarchical model-building approaches.

Social science researchers often compare multilevel models (MLMs) differing in fixed and/or random effects using, for instance, information criteria (Hamaker et al., 2011; Pu & Niu, 2006) or likelihood ratio tests (Stram & Lee, 1994; Vong et al., 2012). In this MLM comparison context, researchers have also had continuing interest in using differences in R-squared (ΔR^2) measures as a way to communicate effect size and importance of included terms (e.g., American Psychological Association, 2008; Bickel, 2007; Dedrick et al., 2009; Edwards et al., 2008; LaHuis et al., 2014; Nakagawa & Schielzeth, 2013; Peugh, 2010; Xu, 2003). However, there has been confusion regarding which of the available MLM R^2 measures should be used in comparing different types of MLMs. Methodologists have sought to inform this practice with several recent simulation studies on the use of R^2 in MLM comparisons (Jaeger, Edwards, Das,

& Sen, 2017; Jaeger, Edwards, & Gurka, 2019; LaHuis et al., 2014; Orelien & Edwards, 2008; Wang & Schaalje, 2009). However, the procedures used in these studies had one or more of the following three limitations (listed here but explained later):

1. Using a measure that, by definition, cannot reflect the model manipulation because the term(s) added to the model do not affect the components considered to be *explained* variance.
2. Using a measure that, by definition, cannot reflect the model manipulation because the term(s) added to the model do not affect the components considered to be *unexplained* variance.
3. Using only measures reflecting differences in *total* variance explained, while neglecting or incorrectly characterizing measures reflecting differences in *level-specific* variance explained.

These limitations led recent simulation studies to provide some misleading and incomplete recommendations for practice, such as the recommendation to avoid altogether a measure that should not be used for certain purposes but could instead be used for other purposes (as we will show). Here we make the following changes to previous procedures. First, rather than the common (but, as we explain later, unfruitful) previous practice of seeking a single, one-size-fits-all R^2 measure relevant to all kinds of MLM comparisons, we define a cohesive *set* of measures for model comparison; this set is sufficiently general to allow researchers to convey effect size for any MLM comparison that has been investigated in previous simulations on ΔR^2 for MLM. Only 4 out of the 12 ΔR^2 measure(s) in this set have been used in previous MLM comparisons. Furthermore, we provide researchers with a clear, step-by-step procedure for identifying *which* ΔR^2 measure(s) within our set reflect differences in explained variance relevant to *which* specific term(s) that could be included/excluded in a given model comparison. Additionally, we explain and concretely demonstrate how the three limitations (listed above) of previous simulation studies on MLM ΔR^2 arose from not using our step-by-step procedures and from not having access to and information about certain measures within our set.

The remainder of this paper proceeds as follows. First, we review the multilevel data model. Second, we briefly review a general framework of MLM R^2 measures that was originally developed for evaluating a single hypothesized model in isolation. Third, we show how this framework can be adapted to, and interpreted in, the context of model comparison. Fourth, we describe step-by-step procedures for identifying ΔR^2 measures that by definition reflect specific kinds of explained variance differences between models. Fifth, we demonstrate each of the three limitations (listed above) of prior MLM ΔR^2 simulation studies, and show how they can be avoided by using our framework and procedures. Sixth, we demonstrate our recommended procedure for using ΔR^2 in an illustrative application comparing six fitted MLMs using software newly developed for this purpose. We conclude with discussion of implications and recommendations for practice and note several extensions and avenues for future research.

Review of the multilevel data model

To begin, we review the multilevel data model. The observation-level (level-1) model is:

$$y_{ij} = \beta_{0j} + \sum_{p=1}^{P} \beta_{pj} v_{pij} + e_{ij} \qquad (1)$$

$$e_{ij} \sim N(0, \sigma^2)$$

Here, we are modeling some continuous outcome y_{ij} for observation i nested within cluster j. The level-1 residual e_{ij} is normally distributed with variance σ^2. The intercept and slopes may vary by cluster, where β_{0j} is cluster j's intercept and β_{pj} is cluster j's slope for the pth level-1 predictor v_{pij} ($p = 1, \ldots, P$). The cluster-level (level-2) model is given as:

$$\beta_{0j} = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{qj} + u_{0j}$$

$$\beta_{pj} = \gamma_{p0} + \sum_{q=1}^{Q} \gamma_{pq} z_{qj} + u_{pj} \qquad (2)$$

$$\mathbf{u}_j \sim MVN(0, \mathbf{T})$$

The cluster-specific intercept (β_{0j}) is composed of a fixed component (γ_{00}), plus the sum of all Q level-2 predictors (z_{qj}'s; $q = 1, \ldots, Q$) multiplied by their slopes (γ_{0q}'s), plus the cluster-specific intercept residual (u_{0j}). Similarly, the pth cluster-specific slope (β_{pj}) is composed of a fixed component (γ_{p0}), plus the sum of all Q level-2 predictors (z_{qj}'s) multiplied by their slopes (γ_{pq}'s), plus the pth cluster-specific slope residual (u_{pj}). Each γ_{pq} thus denotes a cross-level interaction between z_{qj} and v_{pij}. Certain terms in the level-2 model may be set to 0; for instance, if a level-2 predictor z_{qj} is not used to model a cluster-specific intercept or slope, the corresponding γ_{0q} or γ_{pq} would be 0. Level-2 residuals in $\mathbf{u}_j$, a $(P+1) \times 1$ vector containing u_{0j} and P u_{pj}'s, are multivariate normally distributed with covariance matrix $\mathbf{T}$. For a fixed intercept model, the u_{0j}'s are all set to 0; similarly, a fixed slope for the pth level-1 predictor is obtained by setting the u_{pj}'s to 0.

For simplicity of presentation and to facilitate later computation, the above level-1 and level-2 models can be combined into a single reduced-form expression, separating out the fixed components, the γ's, and the random components, the u's, like so:

$$y_{ij} = \left(\gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{qj} + \sum_{p=1}^{P} \gamma_{p0} v_{pij} + \sum_{p=1}^{P} v_{pij} \sum_{q=1}^{Q} \gamma_{pq} z_{qj} \right) \\ + \left(u_{0j} + \sum_{p=1}^{P} u_{pj} v_{pij} \right) + e_{ij}$$

$$(3)$$

which can be further simplified into vector form as:

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\gamma} + \mathbf{w}'_{ij}\mathbf{u}_j + e_{ij}. \quad (4)$$

The first part of the Equation (4) model reflects the fixed portion, with $\mathbf{x}'_{ij}$ denoting a row vector of 1 (for the intercept) and all (level-1 or level-2) predictors, and $\boldsymbol{\gamma}$ denoting a column vector of corresponding fixed effects (i.e., fixed components of the intercept and slopes). Cross-level interaction terms would also go into $\mathbf{x}'_{ij}$, with their corresponding fixed component going into $\boldsymbol{\gamma}$. The second part of the Equation (4) model denotes the random portion, with $\mathbf{w}'_{ij}$ denoting a row vector of 1 (for the intercept) and all (level-1) predictors and $\mathbf{u}_j$ denoting a column vector of corresponding random effect residuals (some of which would be set to 0 if a given slope or the intercept is fixed).

Hereafter, we assume that all level-1 predictors in the model are cluster-mean-centered for the following two reasons.[1] First, this allows researchers to substantively interpret slopes as reflecting purely within-cluster effects or between-cluster effects rather than an "uninterpretable blend" of the two (Cronbach, 1976), as has been widely recommended (e.g., Algina & Swaminathan, 2011; Curran et al. 2012; Enders & Tofighi, 2007; Hofmann & Gavin, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Second, this facilitates the decomposition of proportions of outcome variance explained into separate within-cluster and between-cluster portions (Rights & Sterba, 2019). The cluster-mean-centered model can be given as:

$$y_{ij} = \mathbf{x}^{w\prime}_{ij}\boldsymbol{\gamma}^w + \mathbf{x}^{b\prime}_j\boldsymbol{\gamma}^b + \mathbf{w}'_{ij}\mathbf{u}_j + e_{ij}. \quad (5)$$

In this re-expression, the fixed portion is broken down into within-cluster and between-cluster parts, with $\mathbf{x}^{w\prime}_{ij}$ denoting a vector of level-1 predictors (also including any cross-level interactions), $\mathbf{x}^{b\prime}_j$ denoting a vector of 1 (for the intercept) and level-2 predictors, and $\boldsymbol{\gamma}^w$ and $\boldsymbol{\gamma}^b$ denoting the corresponding within-cluster and between-cluster fixed effects, respectively.

Review of a general framework of MLM R^2 measures

Rights and Sterba (2019) developed a general framework of MLM R^2 measures that subsumed previous commonly used MLM R^2 measures (from: Aguinis & Culpepper, 2015; Hox, 2010; Johnson, 2014; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012; Vonesh & Chinchilli, 1997; Xu, 2003) as special cases, identified redundancies among previous measures, and provided new measures to fill particular

substantive needs. First, we briefly review this framework. Then we describe how to adapt it to the context of model comparison.

Full partitioning of variance

The general MLM R^2 framework employed the following partitioning of model-implied total outcome variance.

$$
\begin{aligned}
\textit{model-implied total outcome variance} \\
= \text{var}(\mathbf{x}^{w\prime}_{ij}\boldsymbol{\gamma}^w + \mathbf{x}^{b\prime}_j\boldsymbol{\gamma}^b + \mathbf{w}'_{ij}\mathbf{u}_j + e_{ij}) \\
= \boldsymbol{\gamma}^{w\prime}\boldsymbol{\Phi}^w\boldsymbol{\gamma}^w + \boldsymbol{\gamma}^{b\prime}\boldsymbol{\Phi}^b\boldsymbol{\gamma}^b + tr(\mathbf{T}\boldsymbol{\Sigma}) + \tau_{00} + \sigma^2
\end{aligned} \quad (6)
$$

This is a more complete partitioning than used previously for developing MLM R^2 measures (e.g., by Snijders & Bosker, 2012; Nakagawa & Schielzeth, 2013; or Johnson, 2014), and hence it provided more possibilities and flexibility in defining measures. Here, $\boldsymbol{\Phi}^w$ denotes the covariance matrix of $\mathbf{x}^w_{ij}$, and $\boldsymbol{\Phi}^b$ denotes the covariance matrix of $\mathbf{x}^b_j$. Also, $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{w}'_{ij}$. The level-2 random intercept variance is given as τ_{00}. Equation (6) represents the sum of five distinct variances, given in Equations (7)–(11), each of which reflect variance attributable to one of five *sources*. Below we supply shorthand symbols for the first four sources ("f_1," "f_2," "v", "m").[2]

$\boldsymbol{\gamma}^{w\prime}\boldsymbol{\Phi}^w\boldsymbol{\gamma}^w$ = variance attributable to *level-1 predictors via fixed components of slopes*
    shorthand : variance attributable to "f_1"
$$(7)$$

$\boldsymbol{\gamma}^{b\prime}\boldsymbol{\Phi}^b\boldsymbol{\gamma}^b$ = variance attributable to *level-2 predictors via fixed components of slopes*
    shorthand : variance attributable to "f_2"
$$(8)$$

$tr(\mathbf{T}\boldsymbol{\Sigma})$ = variance attributable to *level-1 predictors via random slope variation/covariation*
    shorthand : variance attributable to "v"
$$(9)$$

τ_{00} = variance attributable to *cluster-specific outcome means via random intercept variation*
    shorthand : variance attributable to "m"
$$(10)$$

σ^2 = variance attributable to *level-1 residuals* $\quad (11)$

The model-implied *within-cluster* outcome variance is given as:

---

[1]Nonetheless we address ΔR^2 computation for non-cluster-mean-centered MLMs in the discussion.

[2]Note that the level-2 random intercept-slope covariances drop out of the model-implied total outcome variance expression when level-1 predictors are cluster-mean-centered (see Rights & Sterba [2019] Appendix A Equation (A7). For non-cluster-mean-centered models (discussed later), the intercept-slope covariances are involved in the computation of variance attributable to source m (see Rights & Sterba [2019] Table 5).

$$\text{model-implied within-cluster outcome variance} = \boldsymbol{\gamma}^{w\prime}\boldsymbol{\Phi}^w\boldsymbol{\gamma}^w + tr(\mathbf{T}\boldsymbol{\Sigma}) + \sigma^2 \tag{12}$$

Note that this expression includes variance attributable to sources "f_1" and "v."

Finally, the model-implied between-cluster variance is given as:

$$\text{model-implied between-cluster outcome variance} = \boldsymbol{\gamma}^{b\prime}\boldsymbol{\Phi}^b\boldsymbol{\gamma}^b + \tau_{00} \tag{13}$$

Note that this expression includes variance attributable to sources "f_2" and "m."

Throughout this paper, we make a general distinction between *outcome variance* versus *residual variance*. Level-1, or within-cluster, *outcome* variance refers to variance in the outcome (y_{ij}) across level-1 units within level-2 units (i.e., $\text{var}_{i|j}(y_{ij})$), which can include variance attributable to each of f_1, v, and level-1 residuals, as shown in Equation (12). Level-1 *residual* variance, in contrast, solely refers to variance in the level-1 residuals, i.e., σ^2. Similarly, level-2, or between-cluster, *outcome* variance refers to variance in the outcome across level-2 units (i.e., $\text{var}_j(y_{ij})$), which can include variance attributable to each of f_2 and m, as shown in Equation (13). Level-2 *residual* variance, in contrast, refers to either variance in the random intercept and/or variance in a random slope (i.e., diagonal elements of $\mathbf{T}$).[3]

Defining total and level-specific MLM R^2 measures

An R^2 measure can be generically defined in the population as

$$R^2 = \frac{explained\ variance}{outcome\ variance}. \tag{14}$$

Hence, defining an R^2 for MLM involves two considerations: 1) what *outcome variance* is of interest (i.e., "what goes into the denominator?"), and 2) what sources contribute to *explained variance* (i.e., "what goes into the numerator?"). Rights and Sterba's (2019) framework of R^2 measures for MLM is reviewed in Appendix A, wherein each measure's *subscript* refers to what it considers outcome variance (in the denominator) and each measure's *superscript* refers to

what it considers sources of explained variance (in the numerator).

Regarding "what goes into the denominator?" of an MLM R^2, there are three possibilities: total vs. within-cluster. vs. between-cluster outcome variance (Equations (6) vs. (12) vs. (13)). Hence, the framework in Appendix A distinguishes among: total R^2 measures, (having total variance in the denominator as denoted with an "t" subscript), within-cluster R^2 measures (having within-cluster variance in the denominator as denoted with a "w" subscript), and between-cluster R^2 measures (having between-cluster variance in the denominator as denoted with a "b" subscript).

Regarding "what goes into the numerator?" of a total MLM R^2 measure, Rights and Sterba's (2019) framework reviewed in Appendix A shows that variance attributable to the superscripted sources "f_1" "f_2," "v" and/or "m" (from Equations (7)-(10)) can be used singly to create a *single-source measure* (quantifying variance explained by a single source, i.e., either "f_1", "f_2", "v", or "m") or in combination to create a *combination-source measure* (quantifying variance explained by some combination of "f_1", "f_2", "v", and "m"). In the numerator of a within-cluster MLM R^2 measure, superscripted relevant sources "f_1" or "v" can be used singly or in combination, and in the numerator of a between-cluster MLM R^2 measure, superscripted relevant sources "f_2" or "m" are used singly.[4]

To facilitate interpreting all measures in the framework together in an integrated manner, Figure 1 graphically depicts all of these MLM R^2s for a hypothetical example. The left bar contains total R^2s, the middle bar contains within-cluster R^2s, and the right bar contains between-cluster R^2s. Each shaded segment in a bar represents a single-source R^2 measure with the corresponding symbol from Appendix A superimposed on each segment. The legend also indicates which specific source is represented by each segment. Combination-source R^2 measures from Appendix A are visualized by combining shaded segments in a given bar chart. Note that the blank (white) portion in a bar chart represents the proportion of scaled level-1 residual variance, which is not included as a source of explained variance in any measure.

---

[3]This implies, perhaps counterintuitively, that one type of *level-2 residual* variance (the random slope variance) only contributes to *level-1 outcome* variance, not *level-2 outcome* variance. While one might expect that the level-2 residual variance would be a subset of the level-2 outcome variance, in fact it is not because, although a random slope residual varies only between cluster, the *product* of a cluster-mean-centered level-1 variable and a random slope residual varies only within-cluster and thus contributes only to level-1 outcome variance (as explained in Rights & Sterba, 2019).

[4]Note that it is unnecessary to compute a combined-source between-cluster measure, since the two sources "f_2" and "m" together account for all of the between-cluster variance. That is, by definition, f_2 reflects between-cluster variance accounted for by level-2 predictors and m reflects between-cluster variance that is not accounted for by level-2 predictors.
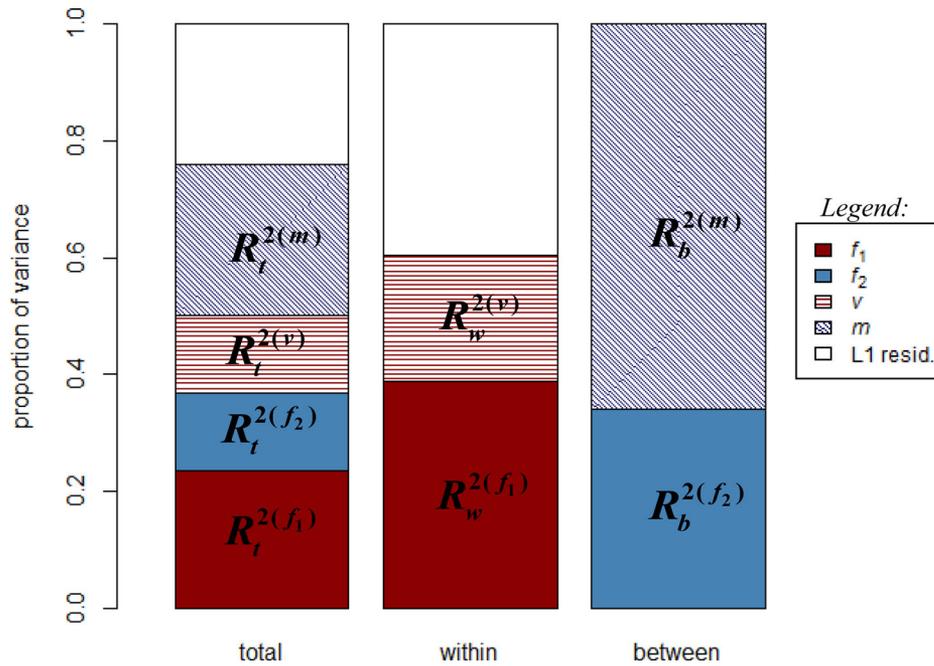
Figure 1. Review: Graphical depiction of multilevel model R^2 measures in Rights and Sterba's (2019) R^2 framework for a single fitted multilevel model. Here hypothetical model results are shown (see Appendix A for more details).

Notes. The left, middle, and right bar charts depict the decomposition of proportions of scaled total, within-cluster, or between-cluster variance, respectively. As such, the left bar chart contains total R^2 measures. The middle bar chart contains within-cluster R^2 measures. The right bar chart contains between-cluster R^2s. Each shaded segment in a given bar represents a *single-source* R^2 measure and the corresponding symbol from Appendix A is superimposed on each segment. The legend also indicates which specific source is represented by each segment; shorthand symbols for sources "f_1", "f_2", "v", and "m" were defined in the manuscript. *Combination-source* measures from Appendix A are visualized by combining shaded segments within a given bar. Total *single-source* measures are: $R_t^{2(f_1)}$, $R_t^{2(f_2)}$, $R_t^{2(v)}$, and $R_t^{2(m)}$. Total *combination-source* measures are: $R_t^{2(f)} = R_t^{2(f_1)} + R_t^{2(f_2)}$, $R_t^{2(fv)} = R_t^{2(f_1)} + R_t^{2(f_2)} + R_t^{2(v)}$, and $R_t^{2(fvm)} = R_t^{2(f_1)} + R_t^{2(f_2)} + R_t^{2(v)} + R_t^{2(m)}$. Within-cluster *single-source* measures are: $R_w^{2(f_1)}$ and $R_w^{2(v)}$. Within-cluster *combination-source* measures are: $R_w^{2(f_1 v)} = R_w^{2(f_1)} + R_w^{2(v)}$. Between-cluster *single-source* measures are: $R_b^{2(f_2)}$ and $R_b^{2(m)}$. Note that the blank (white) portion in a bar chart represents the scaled level-1 residual variance, which is not included as a source of explained variance in any measure.

Applying the MLM R^2 framework to the context of model comparison

Many methods with various strengths and weaknesses exist for helping researchers compare and rank competing MLMs to aid in model building, such as information criteria or likelihood ratio tests (for reviews, see Dimova, Markatou, & Talal, 2011; Dunson, 2008; Fan & Li, 2012; Müller, Scealy, & Welsh, 2013; Ryoo, 2011). MLM ΔR^2s can complement these methods by communicating the effect size—on a familiar proportion metric ranging from 0 to 1—associated with adding terms to the MLM.

In recent methodological simulation studies (Jaeger et al., 2017; LaHuis et al., 2014; Orelien & Edwards, 2008; Wang & Schaalje, 2009), several R^2 measures were used in the context of MLM comparisons. In these simulation studies, each R^2 measure was computed for two candidate models, with their difference here denoted ΔR^2. For a given pair of candidate MLMs, these simulations evaluated each such ΔR^2

measure's ability to reflect term(s) present in the population model (by seeing whether each R^2 measure increases when these term(s) were added to the fitted model). Typically, these simulation studies' ultimate goal was explicitly or implicitly to find a single one-size-fits-all MLM ΔR^2 measure that could be interpreted on its own and would be sensitive to both fixed effects and random effects misspecification. However, no measure was found to fulfill the omnibus goal of serving as a one-size-fits-all MLM ΔR^2 measure of effect size.[5] Generalizing these results, this led

[5]We do not consider either Edwards et al.'s (2008) R-squared-beta measure (also implemented by Jaeger et al., 2017) or Jaeger et al.'s (2019) R-squared-Sigma measure as an omnibus effect size measure for MLM comparisons that include differences in both fixed and random effects, and these measures' authors agree with us (see Edwards et al., [2008, p. 6138, 6154] and Jaeger et al. [2019, p. 169]). In particular, R-squared-beta was defined for "comparing nested mean models with the same covariance structure" and recommended "for assessing fixed effects in the linear mixed model" (Edwards et al., [2008] p. 6138, 6154). Subsequently, R-squared-Sigma was defined specifically to "conduct covariance model selection" and recommended for use in comparing models "differing only by covariance specification" (Jaeger et al., 2019, p.

some to wonder whether ΔR^2 measures simply have limited use in MLM comparisons altogether (e.g., Orelien & Edwards, 2008) and also led to a variety of specific points of confusion addressed later in our section on *Limitations of procedures previously used for comparing MLM R^2*.

Our goal here is fundamentally different in that we are not seeking a one-size-fits-all ΔR^2 measure for MLM comparisons that is sensitive to inclusion/exclusion of any kind of fixed effect (including a fixed intercept or fixed slopes of level-1 predictors, level-2 predictors, and cross-level interactions) or any kind of random effect (including random intercepts and random slopes). Instead, we are seeking a *match* between a given *kind* of manipulation (e.g., a kind of term omitted from a more-parsimonious Model A but included in a more-complex Model B) and a single-source ΔR^2 measure from our framework that is sensitive to this manipulation, with the understanding that no measure will be sensitive to all kinds of manipulations. Rather, the single-source ΔR^2 measures can be used as an *integrative set* which, taken together, is sensitive to each kind of manipulation. Our approach was foreshadowed by Kramer (2005) and Edwards et al. (2008, p. 6138) who concluded that "Different problems necessarily emphasize the importance of different parts of a model—this is a fundamental component to modeling a process and cannot be resolved mathematically. Thus, there can be no general definition of R^2 for mixed models that will cover every model." Our role here is to provide a clear, straightforward decision-making procedure that leads researchers step-by-step to identify which MLM ΔR^2 measure can be used to reflect which kind of meaningful difference between models. Before describing this procedure, we begin by identifying and filling in some gaps in critical background information that were left by prior methodological studies of ΔR^2 for MLM.

164, 178). However, neither R-squared-beta nor R-squared-Sigma correspond (even approximately) with the population-generating proportion of variance attributable to any source (f_1, f_2, f, v, m) or combination of sources. (To see this, compare Jaeger et al.'s [2017] simulation results in their Table 1 column 1 cell 4, and the difference between cell 4 and cell 5, each to their corresponding population generating quantities, which are listed in our Limitation #1 section. Also compare Jaeger et al.'s [2019] simulation results in their Figure 1 [panel labeled "Model"] with the population generating values of $R_t^{2(f_1)} = .57$, $R_t^{2(f_2)} = 0$, $R_t^{2(v)} = .01$, and $R_t^{2(m)} = .10$; these generating values were computed using formulae outlined in Rights & Sterba, 2019). In contrast, in the present paper, the ΔR^2 we discuss are tools to measure effect size, interpretable as estimates of these meaningful population quantities (following Kelley & Preacher, 2012).

Full delineation of a set of ΔR^2 measures

Prior simulation studies on ΔR^2 in MLM acknowledged only a few of the possible ΔR^2 that could be used as effect sizes for MLM comparisons. The first gap we fill is to broaden these possibilities by delineating and defining, in Table 1 Columns 1 and 2, ΔR^2 measures that could be used in comparing MLMs; these are organized into total vs. within-cluster vs. between-cluster measures. The Table 1 ΔR^2 measures are obtained by computing R^2 measures in the Appendix A framework (see Rights & Sterba, 2019) for both candidate models (A and B), and then taking their difference.[6] We clarify in Table 1 Column 3 that prior methodological studies evaluating ΔR^2 for MLM comparisons used measures equivalent in the population to 4 out of the 12 ΔR^2 measures listed (i.e., measures from Johnson's [2014] extension of Nakagawa and Schielzeth [2013], from Snijders & Bosker [2012], from Xu [2003], from Vonesh and Chinchilli [1997], and from Raudenbush and Bryk [2002]). As such, 8 out of 12 ΔR^2 measures in Table 1 have not been recognized in the methodological literature for the purposes of MLM comparison.

Note that we exclude from Table 1 two existing ΔR^2 measures for the following reasons. First, though we include in Table 1 measures equivalent in the population to Raudenbush and Bryk's (2002, p. 79) proportion reduction in residual variance at level-1 computed using a random-intercept-only null model for each candidate model A and B (here termed $\Delta R_w^{2(f_1 v)}$) and Raudenbush and Bryk's (2002, p. 74) proportion reduction in random intercept variance at level-2 computed using a random-intercept-only null model for each candidate model A and B (here termed $\Delta R_b^{2(f_2)}$), we do not include a measure corresponding to Raudenbush and Bryk's (2002, p. 85) proportion reduction in random *slope* variance at level-2 because it is not interpretable as either a total, within-cluster, or between-cluster ΔR^2. Raudenbush & Bryk's (2002) measures have been called "pseudo-R^2s" that each assess the proportion reduction in residual variance (i.e., the proportion reduction in each of level-1 residual variance, level-2 random intercept variance, or level-2 random slope

---

[6]In Table 1 computations, the denominator of the R^2 for Model A uses a model-implied expression for the outcome variance obtained from Model A and the denominator of the R^2 for Model B uses a model-implied expression for the same outcome variance, obtained from Model B. This same approach has been used in prior methodological studies when computing ΔR^2 (i.e., studies employing measures from Johnson, 2014; Nakagawa & Schielzeth, 2013; Snijders & Bosker, 2012 in the context of model selection). Furthermore, in simulated checks, this model-implied outcome variance was virtually identical regardless of the fitted model (i.e., A or B).

Table 1. ΔR^2 measures for multilevel model (MLM) comparisons: What they quantify and what types of model differences they can reflect.

ΔR^2 Measure:	Definition (Interpretation):	Has measure been used previously in simulation studies of ΔR^2 for MLM?†	By definition, this ΔR^2 is designed only to reflect these terms:	Our recommendation for usage:
Total MLM ΔR^2 measures				
$\Delta R_t^{2(f_1)} = R_{t(B)}^{2(f_1)} - R_{t(A)}^{2(f_1)}$	Difference in the proportion of total outcome variance explained by level-1 predictors via fixed components of slopes	No	• fixed component of level-1 predictor's slope • cross-level interaction	Use this single-source measure only for quantifying contribution of level-1 predictor(s) or cross-level interaction(s) via fixed component(s) of slope(s)
$\Delta R_t^{2(f_2)} = R_{t(B)}^{2(f_2)} - R_{t(A)}^{2(f_2)}$	Difference in the proportion of total outcome variance explained by level-2 predictors via fixed components of slopes	No	• level-2 predictor's slope	Use this single-source measure only for quantifying contribution of level-2 predictor(s)
$\Delta R_t^{2(v)} = R_{t(B)}^{2(v)} - R_{t(A)}^{2(v)}$	Difference in the proportion of total outcome variance explained by level-1 predictors via random slope (co)variation	No	• random slope	Use this single-source measure only for quantifying contribution of level-1 predictor(s) via random slope(s)
$\Delta R_t^{2(m)} = R_{t(B)}^{2(m)} - R_{t(A)}^{2(m)}$	Difference in the proportion of total outcome variance explained by cluster-specific outcome means via random intercept variation	No	• random intercept*	Use this single-source measure only for quantifying contribution of a random intercept
$\Delta R_t^{2(f)} = R_{t(B)}^{2(f)} - R_{t(A)}^{2(f)}$	Difference in the proportion of total outcome variance explained by level-1 and level-2 predictors via fixed components of slopes	Yes[b,c,d] (but prior usage was subject to Limitation 1, described later)	• fixed component of level-1 predictor's slope • cross-level interaction • level-2 predictor's slope	This combination-source measure is an optional supplement to its constituent single-source measures $\Delta R_t^{2(f_1)}$ and $\Delta R_t^{2(f_2)}$, not a standalone measure.
$\Delta R_t^{2(fv)} = R_{t(B)}^{2(fv)} - R_{t(A)}^{2(fv)}$	Difference in the proportion of total outcome variance explained by level-1 and level-2 predictors via fixed slopes and level-1 predictors via random slope (co)variation	No	• fixed component of level-1 predictor's slope • cross-level interaction • random slope • level-2 predictor's slope	This combination-source measure is an optional supplement to its constituent single-source measures $\Delta R_t^{2(f_1)}$, $\Delta R_t^{2(f_2)}$, and $\Delta R_t^{2(v)}$, not a standalone measure.
$\Delta R_t^{2(fvm)} = R_{t(B)}^{2(fvm)} - R_{t(A)}^{2(fvm)}$	Difference in the proportion of total outcome variance explained by level-1 and -2 predictors via fixed slopes, by level-1 predictors via random slope (co)variation, & by cluster-specific outcome means via random intercept variation	Yes[b,c,d] (but prior usage was subject to Limitation 2, described later)	• fixed component of level-1 predictor's slope • cross-level interaction • random slope	This combination-source measure is an optional supplement to its constituent single-source measures $\Delta R_t^{2(f_1)}$, $\Delta R_t^{2(f_2)}$, $\Delta R_t^{2(v)}$, and $\Delta R_t^{2(m)}$, not a standalone measure.
Within-cluster MLM ΔR^2 measures				
$\Delta R_w^{2(f_1)} = R_{w(B)}^{2(f_1)} - R_{w(A)}^{2(f_1)}$	Difference in the proportion of within-cluster outcome variance explained by level-1 predictors via fixed components of slopes	No	• fixed component of level-1 predictor's slope • cross-level interaction	Use this single-source measure only for quantifying contribution of level-1 predictor(s) or cross-level interaction(s) via fixed component(s) of slope(s)
$\Delta R_w^{2(v)} = R_{w(B)}^{2(v)} - R_{w(A)}^{2(v)}$	Difference in the proportion of within-cluster outcome variance explained by level-1 predictors via random slope (co)variation	No	• random slope	Use this single-source measure only for quantifying contribution of level-1 predictors via random slope(s)

(Continued)

Table 1. Continued.

ΔR² Measure:	Definition (Interpretation):	Has measure been used previously in simulation studies of ΔR² for MLM?[†]	By definition, this ΔR² is designed only to reflect these terms:	Our recommendation for usage:
$\Delta R_w^{2(f,v)} = R_{w(B)}^{2(f,v)} - R_{w(A)}^{2(f,v)}$	Difference in the proportion of within-cluster outcome variance explained by level-1 predictors via fixed components of slopes and random slope (co)variation	Yes[a] (but prior usage was subject to Limitation 3, described later)	• fixed component of level-1 predictor's slope • random slope	This combination-source measure is an optional supplement to its constituent single-source measures $\Delta R_w^{2(f_1)}$ and $\Delta R_w^{2(v)}$, not a standalone measure.
Between-cluster MLM ΔR² measures				
$\Delta R_b^{2(f_2)} = R_{b(B)}^{2(f_2)} - R_{b(A)}^{2(f_2)}$	Difference in the proportion of between-cluster outcome variance explained by level-2 predictors via fixed components of slopes	Yes[a]	• level-2 predictor's slope	Use this single-source measure only for quantifying contribution of level-2 predictor(s)
$\Delta R_b^{2(m)} = R_{b(B)}^{2(m)} - R_{b(A)}^{2(m)}$	Difference in the proportion of between-cluster outcome variance explained by cluster-specific outcome means via random intercept variation	No	• N/A[‡]	Use this single-source measure as an optional supplement when including level-2 predictors.

Notes: Single-source measure = a measure that quantifies variance explained by only one source (f_1, f_2, v, and m); combination-source measure = a measure that quantifies variance explained by multiple source (some combination of f_1, f_2, v, and m). See Appendix A for a review of definitions of each R² measure when fitting a single model. a = LaHuis et al. (2014); b = Orelien and Edwards (2008); c = Wang and Schaalje (2009); d = Jaeger et al. (2017).

[†]Prior simulation studies used population-equivalent expressions for $\Delta R_t^{2(f)}$ (from Johnson, 2014; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012; Vonesh & Chinchilli, 1997), for $\Delta R_t^{2(f,v,m)}$ (from Johnson, 2014; Vonesh & Chinchilli, 1997; Xu, 2003), for $\Delta R_w^{2(f,v)}$ (from Raudenbush & Bryk, 2002) using a random-intercept-only null model), and for $\Delta R_b^{2(f_2)}$ (from Raudenbush & Bryk, 2002, using a random-intercept-only null model). Derivations underlying the population equivalencies of these expressions can be obtained from Rights and Sterba (2019, Appendix B).

*Here we assume Model A has a fixed intercept and Model B has a random intercept. In the other rows we assume Models A and B both contain a random intercept; this is conventional when fitted models are called MLMs and it also facilitates computation of the between-cluster variance.

[‡]We include this measure for completeness of the decomposition (see, e.g., Figure 2) though it isn't used to detect new terms.

---

variance) after adding term(s) to a MLM (see, e.g., Hoffman, 2015; Hox, 2010; Kreft & de Leeuw, 1998). These pseudo-R^2s differ from our framework of measures in that the pseudo-R^2s consider how much of one model's *residual variance* (the level-1 residual variance, the level-2 random intercept variance, or the level-2 random slope variance) is explained by adding terms, whereas our framework of measures considers how much of the *overall outcome variance* (total, within-cluster, or between-cluster outcome variance) is explained by adding terms. Nonetheless, in the Discussion we demonstrate how our framework of measures can be used to obtain these pseudo-R^2 (including that for the level-2 random slope variance) for any conceivable null/reduced model. Additionally, in Table 1 we do not include Edwards et al.'s (2008) R-squared-beta measure, nor Jaeger et al.'s (2019) R-squared-Sigma measure, because differences in these measures between models[7] have been shown in simulations (Jaeger et al., 2017, 2019) to not correspond with the population-generating proportion of variance attributable to any source (f_1, f_2, f, v, m) or combination of sources (see our Footnote 4). Though such correspondence is not critical if the measure is simply to be used as a selection index for ranking, it is critical for the kind of meaningful interpretability of effect size indices sought in this paper.

Systematic explanation of which ΔR² measures can be used for what purpose

The second piece of background information we supply is a systematic explanation of *which* ΔR^2 can be used to reflect (i.e., quantify the contribution of) *which* terms that are added to yield Model B. Specifically, Column 4 in Table 1 lists the kinds of terms that, when present in the population, will increase Model B's R^2 relative to Model A's R^2, thus corresponding to a positive ΔR^2. Previous usage of these measures has not been consistent with Table 1 Column 4, which led to the problems described in a subsequent section: *Limitations of procedures previously used for comparing MLM R^2.*

Clarification on the use of single-source versus combination-source ΔR²

Inspection of the total measures and within-cluster measures in Table 1 reveals that the addition of any

---

[7]Note that Edwards et al.'s (2008) and Jaeger et al.'s (2019) measures can be computed for Models A and B and their difference taken, or a squared semi-partial version of Edwards et al.'s (2008) and Jaeger et al.'s (2019) measures using an approximate F statistic for a Wald test of the relevant model coefficient can be used.

given kind of term to Model B can be reflected in either a *single-source* measure or *combination-source* measure (defined previously). For instance, the addition to Model B of a fixed component of a level-1 predictor's slope can be reflected by a single-source total measure $\Delta R_t^{2(f_1)}$ as well as by combination-source total measures ($\Delta R_t^{2(f)}$, $\Delta R_t^{2(fv)}$, and/or $\Delta R_t^{2(fvm)}$). Table 1 Column 3 shows that prior methodological studies utilizing ΔR^2 measures typically relied only on combination-source measures, without simultaneously considering their constituent single-source measures. Our recommendation is the opposite (as shown in Table 1 Column 5 and detailed in next section)— single-source ΔR^2 are required to be computed and interpreted, whereas combination-source measures are an optional supplement. Single-source ΔR^2 measures have straightforward and clear interpretations; in contrast, combination-source ΔR^2 measures can be unclear and subject to misinterpretation in MLM comparison when used without consideration of the constituent single-source measures (as illustrated in the later section *Limitations of procedures previously used for comparing MLM R^2*).

Step-by-step procedure for using ΔR^2 measures in MLM model comparison

Having supplied the relevant background information in the previous section, we move on to provide a concrete step-by-step decision procedure for using ΔR^2 measures in MLM model comparisons. This procedure is presented in Table 2 and is explained here. In a subsequent section, this procedure is demonstrated with an illustrative example.

Step 1

Suppose we have two MLMs to compare, denoted Model A and Model B. Step 1 is for the researcher to determine the kind of term(s) that will be included in Model B but excluded from Model A. In Table 2 Row 1, this choice is among five columns; a random intercept, fixed component(s) of level-1 predictor(s)' slopes, fixed component(s) of cross-level interaction(s), random slope(s) of level 1 predictor(s), and/or fixed slope(s) of level-2 predictor(s). For instance, a researcher analyzing students nested within schools might add to some Model A a fixed slope of a school-level predictor (e.g., school size) to form Model B; in this case, the researcher would choose "fixed slope(s) of level-2 predictor(s)" at Step 1.

Several details are important to note regarding Step 1. First, if more than two models are to be compared (as in our upcoming illustrative example), these MLMs can be compared in pairs. Also note that, although in practice a pair of MLMs to be compared is commonly nested (meaning that Model A can be obtained by placing constraints on Model B), the pair of MLMs need not be nested when computing our ΔR^2 measures. For instance, one might wish to compare different sets of predictors to assess which explains the most variance. Nonetheless, in this paper our illustrative examples focus on nested model comparisons, as is most typical of practice.

Another detail to note regarding Step 1 is that adding a cross-level (level-1 × level-2) interaction term (e.g., school size × student delinquency) to Model B falls into the same column as adding a "fixed component of a level-1 predictor's slope" to Model B in Table 2. This is because, under our earlier-stated assumption that the level-1 predictor is cluster-mean-centered, this cross-level interaction term can only explain *level-1 outcome* variance (see derivation in Rights & Sterba, 2019).[8] Such cross-level interactions have often previously been understood as explaining level-2 variance in the sense that they reduce the *level-2 random slope* variance of the level-1 predictor (e.g., Hoffman, 2015; Raudenbush & Bryk, 2002). Though this understanding may initially appear inconsistent with our description, in fact it is entirely consistent for the following reason. Adding a cross-level interaction involving a cluster-mean-centered level-1 predictor leads to a reduction in the level-2 *residual* variance (specifically, level-2 random slope variance), but does not account for level-2 *outcome* variance because the product of a level-2 predictor with a cluster-mean-centered level-1 predictor varies only within-cluster, and hence can only explain *outcome* variance at level-1. Similarly, as mentioned earlier, random slope variability contributes only to within-cluster outcome variance because the product of a cluster-mean-centered level-1 variable and a random slope residual also varies only within-cluster.

An additional detail to note is that, in our examples and description below, all MLMs under consideration are assumed to have at least a random intercept unless an initial comparison is being made between a fixed intercept and random intercept null model. Also observe that Table 2 Row 1 has no separate column for the addition of random effect covariances to Model B because we simply assume researchers will

---

[8]See the Discussion section for details regarding non-cluster-mean-centered level-1 predictors.

Table 2. Recommended step-by-step procedure for using ΔR^2 measures in multilevel model (MLM) comparisons (this procedure can be implemented with either the simultaneous or hierarchical model-building approach described later in the Extensions section.).

	Proceed down column(s) at each step to identify which measure(s) to interpret							
	Random intercept*		Fixed component of a level-1 predictor's slope or a cross-level interaction‡		Random slope		Level-2 predictor's slope	
	Total measure	Between-cluster measure	Total measure	Within-cluster measure	Total measure	Within-cluster measure	Total measure	Between-cluster measure
Step 1: What kind of term(s) are in Model B that were not in Model A?								
Step 2: Are you interested in quantifying the impact of added term(s) overall (if so, choose a total measure) or at a particular level (if so, choose a level-specific measure) or both (choose both)?								
Step 3: Compute target single-source ΔR^2 measure(s) that can reflect the importance of added term(s).	$\Delta R_t^{2(m)}$	N/A†	$\Delta R_t^{2(f_1)}$	$\Delta R_w^{2(f_1)}$	$\Delta R_t^{2(v)}$	$\Delta R_w^{2(v)}$	$\Delta R_t^{2(f_2)}$	$\Delta R_b^{2(f_2)}$
Step 4: Visualize and interpret changes in target single-source measure(s) in the context of the set of all single-source measures in the Table 1 framework.	See Figure 2 example	See Figure 2 example	See Figure 2 example	See Figure 2 example	See Figure 2 example	See Figure 2 example	See Figure 2 example	See Figure 2 example
Step 5: (Optional) Decide whether to compute combined-source measures in Table 1 from their constituent single-source measures.	Option to combine with other single-source total ΔR^2		Option to combine with other single-source total ΔR^2	Option to combine with other single-source level-1 ΔR^2	Option to combine with other single-source total ΔR^2	Option to combine with other single-source level-1 ΔR^2	Option to combine with other single-source total ΔR^2	

Notes: Each measure's superscript indicates its source(s) of explained variance: f_1 = level-1 predictors via fixed components of slopes; f_2 = level-2 predictors via fixed components of slopes; v = level-1 predictors via random slope (co)variation; m = cluster-specific outcome means via random intercept variation. Each measure's subscript indicates what it considers as outcome variance: t = total variance, b = between-cluster variance, or w = within-cluster variance.

‡Note that if a researcher adds a cross-level (L1 × L2) interaction to Model B this would explain variance at only level-1 when level-1 predictors are cluster-mean-centered (see Rights & Sterba, 2019).

*In this column we assume Model A has a fixed intercept and Model B has a random intercept. In the other columns we assume Models A and B both contain a random intercept; this is conventional when fitted models are called MLMs and it also facilitates computation of the between-cluster variance.

†We list this measure as N/A here because in this column we are assuming Model A is a fixed intercept null model and Model B is a random intercept null model, in which case this measure is 1.0.

estimate all covariance terms associated with each added random effect.

A final detail to note regarding Step 1 concerns whether to add a single term at a time to the MLM, or multiple terms at a time. Often researchers prefer to add a single term at a time during model building. This is simplest, but it is not necessary here. Because the single-source measures provided in our framework distinguish among the contributions of the level-1 predictors via the fixed components of slopes, the level-1 predictors via random components of slopes, and the level-2 predictors, terms of these three different kinds could be added simultaneously to yield Model B and the single-source ΔR^2 results for these terms would isolate their respective unique contributions to explained variance. For instance, if a level-1 predictor and a level-2 predictor (e.g., student delinquency and school size) were added simultaneously to yield Model B, their respective *unique* contributions to explained variance via fixed slope components would nonetheless still be quantified because the decomposition of variance used in creating the measures will separate their contributions into orthogonal components. On the other hand, if multiple terms of the *same* kind (i.e., from the same column of Table 2) are added simultaneously to yield Model B—for instance, adding three level-2 predictors at once—then only their *joint* contribution to explained variance could be quantified using the ΔR^2s; identifying each of their unique contributions would require entering these terms one at a time, in separate model comparisons. In any case, researchers adding term(s) of only one kind to yield Model B would choose the corresponding column of Table 2 in Step 1 and proceed to Step 2. In contrast, researchers simultaneously adding terms of multiple kinds to yield Model B would choose the multiple corresponding columns of Table 2 in Step 1 and proceed to Step 2.

Step 2

In the second step we proceed to Row 2 of Table 2. Specifically, within the chosen column(s) from Step 1, researchers must in Step 2 decide whether they are interested in quantifying the impact of an added term *overall* (if so, use a total ΔR^2 measure) or at *a particular level* (if so, use a level-specific ΔR^2 measure), or *both* (if so, use both total and level-specific ΔR^2 measures). Depending on the column, the appropriate level-specific measure in Step 2 of Table 2 is either level-1 specific (within-cluster measure) or level-2 specific (between-cluster measure). Note that, for any type of model comparison, a *total* ΔR^2 for a generic single-source *s* will, by definition, be smaller than its counterpart *level-specific* ΔR^2 for the same generic single-source *s*. This is simply because the total variance is larger than the level-specific variance at any one level. As such, a given source can potentially explain little of the total outcome variance, but a large portion of level-specific variance.

Step 3

The third step is to determine which single-source ΔR^2 measure(s) reflect the type of differences deemed of importance in Step 2. This is done in Row 3 of Table 2. Once appropriate target single-source measure(s) are identified, they can be interpreted using the detailed definitions supplied in Table 1. Recall that each target single-source measure provides a quantitative *effect size difference* between two models on the easily interpretable metric of the proportion of variance explained, and allows assessment of *the degree to which* the two models differ as a result of the particular kind of term added to yield Model B.

Step 4

As a fourth step, it is helpful to next consider one's target single-source ΔR^2 measure(s) from Step 3 in the broader context of the set of all other single-source ΔR^2 measures. Doing so provides the advantage of having a complete summary of the differences in explained variance between Models A and B, allowing juxtaposition of the results from all measures, not just target measures that are responsive to the terms added to yield Model B. This can be particularly useful in seeing how an increase in variance attributable to one's target source can lead to a decrease in variance attributable to another source; a common example of this is a cross-level interaction term leading to an *increase* in variance attributable to predictors via fixed effects (the target source) and a *decrease* in variance attributable to predictors via random slope variation.

Step 4 can be done using graphical visualization, by creating bar charts like in Figure 1 for each model under consideration (our software will do this automatically, as described later). Figure 2 provides an example of how to visualize changes in the suite of MLM R^2s across a series of five fitted models taken from our subsequent illustrative example. Figure 2 will be substantively interpreted and described in the subsequent illustrative example section, however, a
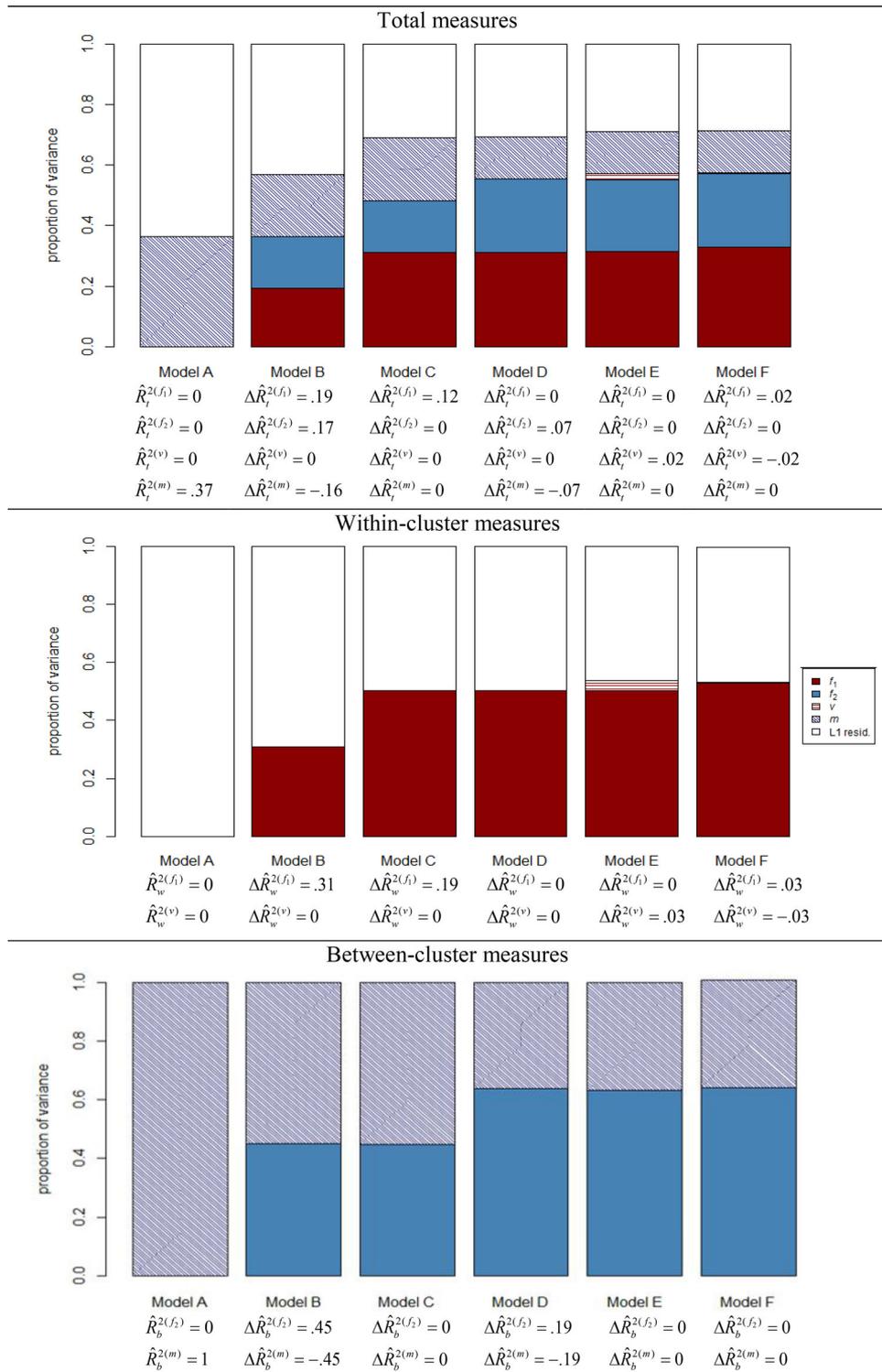
Figure 2. Graphical comparison of multilevel model R^2 measures as they change across the six fitted Models A-F from the illustrative example analysis (using a hierarchical model-building approach).

Note. R^2s computed from Model A are printed below each bar labeled Model A. ΔR^2s between each consecutive pair of models are printed below each bar labeled B-F.

brief overview of the layout of Figure 2 is provided here. Each column of Figure 2 corresponds with a particular model under comparison. Hence, there are six columns in Figure 2 because the illustrative

example involves comparing six models. For a given model (i.e., a given column), the top bar chart depicts single-source total R^2 measures, the middle bar chart (second row) depicts single-source within-cluster R^2

measures, and the bottom bar chart (third row) depicts single-source between-cluster R^2 measures. A researcher can succinctly visually assess change in a single-source R^2 measure of explained variance by viewing how the size of its shaded segment changes across models.

Step 5 (Optional)

Although previous researchers have typically focused on interpreting combination-source ΔR^2 measures to the exclusion of single-source ΔR^2 measures (see Table 1 Column 3), we show later that this can be misleading because it obscures how and which component single-source measures are changing in response to the term(s) added to yield Model B. In contrast, our approach in Steps 1-4 provides researchers with all the information needed to assess changes in variance explained across fitted MLMs without these risks. After assessing changes in target single-source measures in Step 3 and considering these in the context of other single-source ΔR^2 in Step 4, an optional supplemental Step 5 would be to create combination-source ΔR^2 measures listed in Table 1. For a given bar chart in Figure 2, combination-source R^2 measures are created by summing two or more shaded segments. To assess combination-source ΔR^2 measures, simply observe how the sum of the shaded segments changes across models (i.e., across columns of Figure 2).

Software implementation

To aid researchers in computing and visualizing the suite of ΔR^2 measures, we developed an R function, r2MLMcomp. This function is given in the Online Supplementary Material, along with a user guide and example input. To use this R function, researchers input raw data and parameter estimates for two MLMs under comparison at a time. The function then outputs the suite of R^2 measures in Appendix A for each model along with the suite of ΔR^2 measures in Table 1 for comparing the pair of models. A graphical representation of these measures (similar to Figure 2) is also automatically provided.

Limitations of previous procedures for using ΔR^2 measures in MLM comparisons

Our recommended step-by-step procedure (described in the previous section) for comparing an integrated suite of MLM ΔR^2s across fitted models was not implemented in previous methodological simulation studies on this topic. Instead, simulation studies on

using MLM ΔR^2 in model comparisons used procedures with some of the following three limitations. Here we describe each limitation and demonstrate how it affected previous authors' results and conclusions. Then we demonstrate how to overcome each limitation by applying procedures from the previous section.

Limitation 1: Using a measure that, by definition, cannot reflect the model manipulation because the term(s) added to the model do not affect the components considered to be explained variance

Recall that certain R^2 measures are not suited to detect certain types of differences between models, as shown in column 4 of Table 1. When a given addition to Model A can affect portions of *unexplained* variance but cannot affect portions of *explained* variance, the ΔR^2 is incapable of detecting differences between Models A and B. In this instance, there is a mismatch between the model manipulation done and the ΔR^2 used. This general concept can be formalized mathematically in the population as follows. First, recall that the outcome variance is equal to the explained variance plus the unexplained variance. In the context of model comparison, wherein A and B subscripts denote variance obtained from Model A and B, respectively:

$$outcome\ variance = explained_A + unexplained_A \\ = explained_B + unexplained_B. \tag{15}$$

Note that changing the model does not change the variance in the outcome. When an addition to Model A does not affect portions of variance considered explained, this implies that $explained_A = explained_B$. Consequently the ΔR^2 measure will be 0:

$$\begin{aligned} \Delta R^2 &= R^2_{modelB} - R^2_{modelA} \\ &= \frac{explained_B}{outcome\ var.} - \frac{explained_A}{outcome\ var.} \\ &= \frac{explained_A}{outcome\ var.} - \frac{explained_A}{outcome\ var.} \\ &= 0 \end{aligned} \tag{16}$$

Such an addition could still change portions of variance that the ΔR^2 measure counts as unexplained, but does so in compensatory ways such that $unexplained_A = unexplained_B$.

Nonetheless, methodological simulation studies have used ΔR^2 measures in model comparisons wherein the explained variance would not be capable of changing in the population (e.g., Jaeger et al., 2017, 2019), and these studies subsequently critiqued the ΔR^2 measure's inability to detect the term(s) added to

Model A to create Model B. For example,[9] Jaeger, Edwards, Das, & Sen (2017) compared a random slope model (Model B, their generating model) to a random intercept model (Model A, an underspecified model) with the same fixed effects. They found that a measure[10] of $\Delta R_t^{2(f)}$ was unable to detect any difference between Models B and A and then criticized the use of this measure on the grounds that it "cannot distinguish the correct covariance structure" (p. 1096). However, $\Delta R_t^{2(f)}$ is not designed to distinguish the correct covariance structure and should not be used for this purpose in the first place. Adding a random component of a slope does not affect the components considered to be explained variance in $\Delta R_t^{2(f)}$. Rather, adding a random component of a slope affects variance attributable to "v," and "v" contributes only to unexplained variance in $\Delta R_t^{2(f)}$; in this case an increase in "v" from Model A to B will be accompanied by an equivalent decrease in unexplained variance attributable to level-1 residuals, to render outcome variance the same from Model A to B. As shown in Table 1, $\Delta R_t^{2(f)}$ is instead able to assess differences in the proportion of variance explained by predictors via *fixed* components of slopes (i.e., it is designed to reflect the addition of fixed components only). For instance, suppose one were analyzing repeated measures nested within individuals and wanted to compare a Model A with a fixed slope of *age* to a Model B with a random slope of *age* (see footnote 10). For this comparison, $\Delta R_t^{2(f)}$ would be 0, although $\Delta R_t^{2(v)}$ could be quite large if there were a great deal of across-person differences in the effect of *age*.

To illustrate how our step-by-step decision-making procedure in Table 2 would avoid the pitfall of picking a measure that is by definition insensitive to the model manipulation, we re-simulated and then reanalyzed the data from Jaeger, Edwards, Das, & Sen (2017).[11]

Population R^2 values for the full Model B are $R_t^{2(f_1)} = .26$, $R_t^{2(f_2)} = .15$, $R_t^{2(f)} = .41$, $R_t^{2(v)} = .02$, and $R_t^{2(m)} = .38$. Population ΔR^2 values are $\Delta R_t^{2(f_1)} = 0$, $\Delta R_t^{2(f_2)} = 0$, $\Delta R_t^{2(f)} = 0$, $\Delta R_t^{2(v)} = .02$, and $\Delta R_t^{2(m)} = 0$.[12] Using their procedure of seeing how $R_t^{2(f)}$ changes between Models A and B, we can replicate Jaeger et al.'s (2017) finding of no apparent difference between the models ($\hat{\Delta R}_t^{2(f)} = .42 - .42 = 0$). Yet when Jaeger et al. (2017) consider $\Delta R_t^{2(f)}$ in isolation, any model differences pertaining to "v" are masked. Next, we reanalyze these data by instead employing our Table 2 step-by-step procedures. As Step 1, we note that the manipulation being done involves adding a random slope to Model A to form Model B. As Step 2, we identify our interest in quantifying the impact of the added term overall (implying we should target a total measure). As Step 3, we use Table 2 to identify the target single-source measure, $\Delta R_t^{2(v)}$, that indicates correctly that Models A and B do in fact differ in variance explained by predictors via random slope variation, as the across-sample average estimate of $\Delta R_t^{2(v)}$ is .02, matching its population value. Though $\Delta R_t^{2(v)}$ is fairly small here due simply to the generating conditions, if it were more pronounced, it would still be unrecognized using $\Delta R_t^{2(f)}$ in isolation. In Step 4 we visualize and interpret the results from target measure $\Delta R_t^{2(v)}$ in the context of other single-source measures; the associated graph is provided in Figure 3. Rather than letting both the explained and unexplained variance be uninterpretable blends of multiple components (as with Jaeger et al.'s [2017] use of a combination-source measure), here we show each component individually so we can confirm that the components theoretically expected not to change in fact do not change (each of $\hat{\Delta R}_t^{2(f_1)}$, $\hat{\Delta R}_t^{2(f_2)}$, and $\hat{\Delta R}_t^{2(m)}$ are roughly 0).

Limitation 2: Using a measure that, by definition, cannot reflect the model manipulation because the term(s) added to the model do not affect the components considered to be unexplained variance

Limitation 2 is similar to Limitation 1 in that it describes a mismatch between the terms added to Model A to create Model B and the ΔR^2 measure used to quantify the importance of these terms. Under Limitation 1, added terms affected only portions of the unexplained variance, but not the portions of the explained variance; Limitation 2 is, in a sense, the opposite in that the mismatch is caused by adding terms affecting only portions of the *explained* variance, but not portion(s) of the *unexplained* variance

---

[9]What follows is just one example of how Limitation 1 could manifest. Another example occurs anytime a researcher seeks to quantify the contribution of a level-1 predictor added to Model A by erroneously using an ΔR^2 quantifying between-cluster outcome variance explained. Similarly, Limitation 1 would also manifest if a researcher seeks to quantify the contribution of a level-2 predictor added to Model A by erroneously using a ΔR^2 quantifying within-cluster outcome variance explained.

[10]Specifically, Jaeger et al. (2017) used Johnson's (2014) marginal ΔR^2 (an extension of Nakagawa and Schielzeth [2013]) that is equivalent to our $\Delta R_t^{2(f)}$ (see Appendix A and Table 2 of the current paper and see Appendix B of Rights & Sterba, 2019).

[11]Following their procedure, we first fit a MLM with a random intercept, random slope of age, and fixed effects for age, gender, and age × gender using restricted maximum likelihood (REML) to the data from Potthoff and Roy (1964). The Potthoff and Roy (1964) data contain repeated measures on 27 children over four time points, ages 8, 10, 12, and 14. Using the obtained parameter estimates as generating values, we then simulated 10,000 datasets of the original sample size and fit to each a random intercept model with fixed effects of age, gender, and age × gender (Model A) as well as a Model B that added a random slope of age.

[12]We also computed the across-sample averaged estimates for these measures and confirmed that our results correspond with the population values.

Total measures
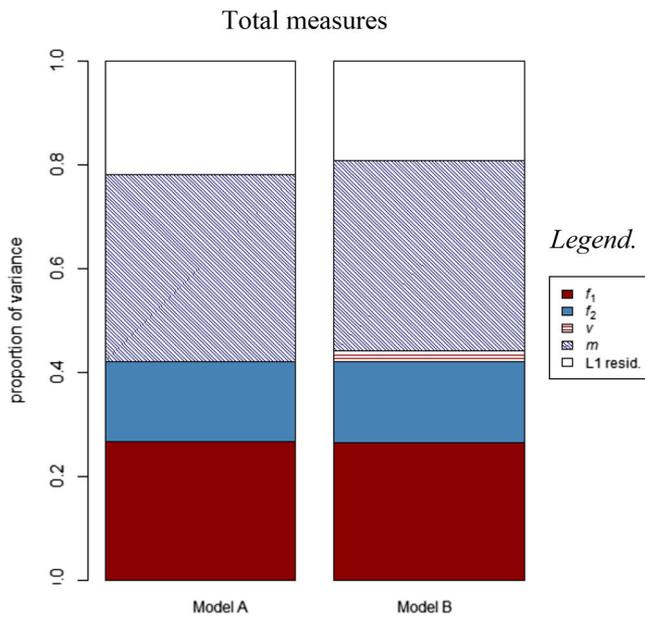
$$\Delta \hat{R}_t^{2(f_1)} = .27 - .27 = 0$$

$$\Delta \hat{R}_t^{2(f_2)} = .16 - .16 = 0$$

$$\Delta \hat{R}_t^{2(f)} = .42 - .42 = 0 \text{ (used in Jaeger et al., 2017)}$$

$$\Delta \hat{R}_t^{2(v)} = .02 - 0 = .02 \text{ (our recommended measure in Table 2)}$$

$$\Delta \hat{R}_t^{2(m)} = .37 - .37 = .00$$

Figure 3. Demonstration of Limitation #1 of previous studies on ΔR^2 for MLM: Use of a measure (here, $\Delta R_t^{2(f)}$) that, by definition, cannot reflect the model manipulation (here, adding a random component of a slope to Model A) because the addition does not affect the components considered to be explained variance in that measure. In contrast, our recommended measure from Table 2 for this model comparison, $\Delta R_t^{2(v)}$, is capable of detecting the addition of the random slope.
Note. Results reported and graphed here are across-sample average results from our replication of the simulation design from Jaeger et al. (2017). Corresponding population values were listed in the text.

(as done in, e.g., Orelien & Edwards, 2008; Jaeger et al., 2017). Illustrating this concept in the population, when a manipulation will not change the unexplained variance (i.e., unexplained$_A$ = unexplained$_B$), this implies that the ΔR^2 will necessarily be 0:

$$\Delta R^2 = R_{modelB}^2 - R_{modelA}^2$$

$$= \frac{\text{explained}_B}{\text{outcome var.}} - \frac{\text{explained}_A}{\text{outcome var.}}$$

$$= \left(1 - \frac{\text{unexplained}_B}{\text{outcome var.}}\right) - \left(1 - \frac{\text{unexplained}_A}{\text{outcome var.}}\right) \quad (17)$$

$$= \left(1 - \frac{\text{unexplained}_A}{\text{outcome var.}}\right) - \left(1 - \frac{\text{unexplained}_A}{\text{outcome var.}}\right)$$

$$= 0$$

Nonetheless, simulations have used ΔR^2 in model comparisons wherein the unexplained variance would not be capable of changing in the population, and have subsequently critiqued these measures' inability to detect the term(s) added to Model A to create Model B (Orelien & Edwards, 2008; Jaeger, Edwards, Das, & Sen, 2017). For instance, Orelien and Edwards (2008) compared a model with just a random intercept and a random slope of a level-1 predictor (Model A) to a model that added two fixed slopes for each of two level-2 predictors (Model B). They found that a measure[13] of $\Delta R_t^{2(fvm)}$ was unable to detect differences between the models. Based on this result, Orelien and Edwards (2008) concluded that "the inadequacy of these R^2 statistics revealed by our simulations put into question their usefulness as a goodness-of-fit (GOF) tool for any mixed model" and then globally recommended that "they should not be used in assessing GOF in the linear mixed model" (p. 1906) because they are "unable to discriminate when important covariates are missing from the model" (p. 1905). Furthermore, this conclusion and global recommendation to avoid the $\Delta R_t^{2(fvm)}$ measure was then restated and reinforced by subsequent authors of other simulations with similar designs (Jaeger, Edwards, Das, & Sen, 2017; Wang & Schaalje, 2009). However, this conclusion and recommendation to avoid using $\Delta R_t^{2(fvm)}$ for all model comparisons is misleading because $\Delta R_t^{2(fvm)}$ *can* reflect the addition of important *level-1* predictors, as shown in column 4 of Table 1. $\Delta R_t^{2(fvm)}$ is simply not suited to detect the specific manipulation done in prior simulations, i.e., the addition of *level-2* predictors, as indicated in column 4 of Table 1. For instance, with repeated observations nested within persons, $\Delta R_t^{2(fvm)}$ could detect the addition of observation-level *age*, but could not detect the addition of person-level *gender*; see footnote 13. $R_t^{2(fvm)}$ counts variance attributable to both f_2 and m as explained, i.e., *all* between-cluster variance is explained. Thus, if a fixed slope of a level-2 predictor is added to an MLM, given the orthogonality of within-cluster and between-cluster components, only variance attributable to f_2 and m could change. Consequently, the unexplained variance would be equivalent between the two models and, regardless of the magnitude of the effect of the level-2 predictor, $\Delta R_t^{2(fvm)}$ would necessarily be 0. In particular, the amount by which variance explained by f_2 would increase would be accompanied by a decrease of equal magnitude in the variance explained by m.

---

[13]Specifically, Orelien and Edwards (2008) used Vonesh and Chinchilli's (1997) conditional ΔR^2 that is equivalent in the population to our $\Delta R_t^{2(fvm)}$ (see Appendix A and Table 1 of the current paper and see Appendix B of Rights & Sterba, 2019).

To illustrate how our step-by-step decision-making procedure in Table 2 would avoid the pitfall of picking a measure that is by definition insensitive to the model manipulation, we re-simulated and reanalyzed the data used in Orelien and Edwards (2008).[14] Population R^2 values of the full Model B were: $R_t^{2(f_1)} = .16$, $R_t^{2(f_2)} = .30$, $R_t^{2(v)} = .005$, $R_t^{2(m)} = .31$, $R_t^{2(fvm)} = .78$. Population ΔR^2 values comparing the full Model B and the reduced Model A excluding the level-2 predictors were: $\Delta R_t^{2(f_1)} = 0$, $\Delta R_t^{2(f_2)} = .30$, $\Delta R_t^{2(v)} = 0$, and $\Delta R_t^{2(m)} = -.30$, yielding $\Delta R_t^{2(fvm)} = 0$.[15] Using Orelien & Edwards' (2008) procedure of examining the combination-source measure $\Delta R_t^{2(fvm)}$ in isolation, we replicate their finding of no apparent difference between Models A and B, as the average estimate of $\Delta R_t^{2(fvm)}$ is 0. Instead using our Table 2 step-by-step procedure, however, we first note that the manipulation being done involves the addition of level-2 predictors (Step 1). Then we identify our interest in quantifying the impact of the added term overall (implying we should use a total measure) (Step 2). Next, we use Table 2 to identify the target single-source measure, $\Delta R_t^{2(f_2)}$, that can reflect the importance of the added terms (Step 3). This target measure $\Delta R_t^{2(f_2)}$ correctly indicates that the two added level-2 predictors explain a sizable portion of the variance (specifically .30, which matches the population value). In Step 4 we visualize and interpret results from our target measure $\Delta R_t^{2(f_2)}$ in the context of other single-source measures, as shown in Figure 4. Then in Step 5 we have the option to compute combination-source measures like $\Delta R_t^{2(fvm)}$, if desired. However, rather than letting the explained variance in this combination-source measure be an uninterpretable blend of four different sources—f_1, f_2, v, and m—as in Orelien and Edwards' (2008) procedure, in Figure 4 we show each single-source component of $\Delta R_t^{2(fvm)}$ individually (i.e., $\Delta R_t^{2(f_1)}$, $\Delta R_t^{2(f_2)}$, $\Delta R_t^{2(v)}$, $\Delta R_t^{2(m)}$), which clarifies that

Total measures
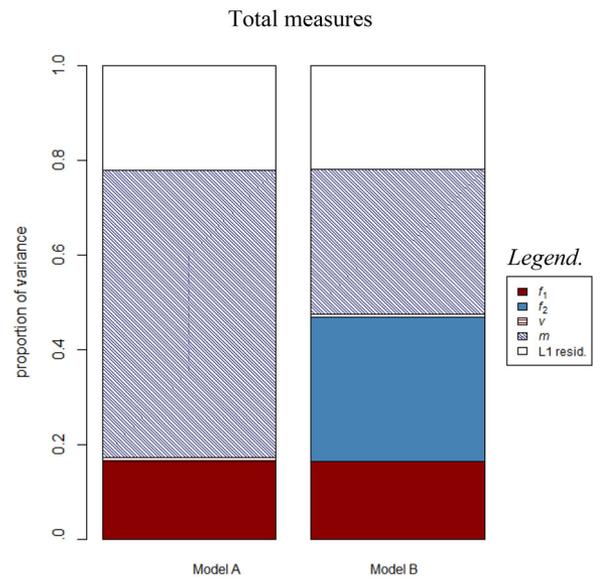
$\Delta \hat{R}_t^{2(f_1)} = .16 - .16 = 0$

$\Delta \hat{R}_t^{2(f_2)} = .30 - 0 = .30$ (our recommended measure in Table 2)

$\Delta \hat{R}_t^{2(v)} = .01 - .01 = 0$

$\Delta \hat{R}_t^{2(m)} = .31 - .61 = -.30$

$\Delta \hat{R}_t^{2(fvm)} = .78 - .78 = 0$ (used in Orelien & Edwards, 2008)

Figure 4. Demonstration of Limitation #2 of previous studies on ΔR^2 for MLM: Use of a measure (here, $\Delta R_t^{2(fvm)}$) that cannot reflect the model manipulation (here, adding fixed slopes of level-2 predictors to Model A) because this addition does not affect the components considered to be unexplained variance in that measure. In contrast, our recommended measure from Table 2 for this model comparison, $\Delta R_t^{2(f_2)}$, is capable of detecting the addition of the level-2 predictors.
Note. Results reported and graphed here are across-sample average results from our replication of the simulation design of Orelien & Edwards (2008). These results match the population values listed in the text.

$\Delta R_t^{2(fvm)} = 0$ simply because the increase in $R_t^{2(f_2)}$ is accompanied by an equal decrease in $R_t^{2(m)}$, as explained above.

Limitation 3: Using only measures reflecting differences in total variance explained, while neglecting or incorrectly characterizing measures reflecting differences in level-specific variance explained

Most simulations addressing ΔR^2 for MLMs have not addressed level-specific ΔR^2 measures. Instead of mentioning that researchers have the option to consider level-specific measures, these studies instead exclusively computed ΔR^2 measures to explain total variance (Jaeger et al., 2017, 2019; Orelien & Edwards, 2008; Wang & Schaalje, 2009). A researcher who is explicitly interested in either the within- or between-cluster variance as a distinct entity would not find changes in total

---

[14]Each of 10,000 generated datasets consisted of 64 persons (level-2 units) each with six observations (level-1 units) at ages 5, 6, 7, 7.25, 7.5, and 7.75 years. The level-2 predictors were treatment and gender, uncorrelated. The level-1 predictor, age, had a random slope. Fixed effects for the intercept, level-1 slope, and two level-2 slopes were 10, 6, 11, and 11, respectively. The random intercept variance was 4, random slope variance was 1, intercept-slope covariance was 1, and level-1 residual variance was 45. Orelien and Edwards (2008) had three separate conditions corresponding with level-1 residual variance of 12, 45, and 250. Here we present the middle value for simplicity. However, we ran the simulation for each condition and our results matched theirs in each case and the same general patterns were found.

[15]Our across-sample average estimates, seen in Figure 4, matched these population values.

variance as informative. Although researchers may often wish to focus on changes in total variance explained, the option of considering level-specific changes can be appealing to researchers in specific applications (see Discussion section), and thus the step-by-step procedures in Table 2 allow researchers the option of considering level-specific measures as well. In particular, Step 2 of Table 2 asks researchers to consider what type of variance they are most interested in explaining: within-cluster variance, between-cluster variance, and/or total variance. Step 3 allows them to choose appropriate level-specific and/or total measures accordingly, and Step 4 additionally allows them to easily visually assess changes in total and level-specific measures simultaneously while focusing interpretation on whichever is of greatest substantive interest.

Though rare, when level-specific measures have been included in simulations on MLM ΔR^2, these level-specific measures have been incorrectly characterized and there have been errors in computing their population values. In particular, LaHuis et al. (2014) stated that they computed one within-cluster measure in their simulation—that of Raudenbush and Bryk (2002, p. 79)—and compared its value for each of two competing random slope models, A and B. The population quantity for the difference between these measures is denoted $\Delta R_w^{2(f_1 v)}$ in Table 1, as evidenced in the derivation provided below in Equation (18):

random slope variation as explained variance, and then erroneously computed its population value as $\Delta R_w^{2(f_1)}$ even though in each sample they computed an estimate of $\Delta \hat{R}_w^{2(f_1 v)}$, which rendered their simulation results pertaining to this measure uninterpretable. Researchers will be able to avoid such misunderstandings in the future by using the definitions of each measure provided in Table 1.

In Equation (18), σ_{RInull}^2 is—as in Raudenbush and Bryk (2002)—the level-1 residual variance from a random-intercept-only null model, i.e., the within-cluster outcome variance. Equation (18) shows that computing the difference between Models A and B in Raudenbush and Bryk's (2002, p. 79) measure reflects differences in the proportion of within-cluster variance explained by level-1 predictors by not only fixed components of slopes (f_1) but also random components of slopes (v).[16]

Illustrative example

In this section, we illustrate our step-by-step procedure for using ΔR^2 measures in MLM comparisons. Our illustration is based on a didactic example from a popular MLM textbook (Hox, 2010). Although the traditional presentation of results focuses on point estimates and standard errors for the final (full) MLM, here we show that inspecting the suite of ΔR^2

$$
\begin{aligned}
&\frac{\sigma_{RInull}^2 - \sigma_B^2}{\sigma_{RInull}^2} - \frac{\sigma_{RInull}^2 - \sigma_A^2}{\sigma_{RInull}^2} \\
&= \frac{(\sigma_{RInull}^2 - \sigma_B^2) - (\sigma_{RInull}^2 - \sigma_A^2)}{\sigma_{RInull}^2} \\
&= \frac{\sigma_A^2 - \sigma_B^2}{\sigma_{RInull}^2} \\
&= \frac{\sigma_A^2 - \sigma_B^2}{\text{within-cluster outcome variance}} \\
&= \frac{\sigma_A^2}{\text{within-cluster outcome variance}} - \frac{\sigma_B^2}{\text{within-cluster outcome variance}} \\
&= \left(1 - \frac{\gamma_{(A)}^{w\,\prime} \mathbf{\Phi}^w \gamma_{(A)}^w + tr(\mathbf{T\Sigma})_{(A)}}{\text{within-cluster outcome variance}}\right) - \left(1 - \frac{\gamma_{(B)}^{w\,\prime} \mathbf{\Phi}^w \gamma_{(B)}^w + tr(\mathbf{T\Sigma})_{(B)}}{\text{within-cluster outcome variance}}\right) \\
&= \left(1 - R_{w(A)}^{2(f_1 v)}\right) - \left(1 - R_{w(B)}^{2(f_1 v)}\right) \\
&= R_{w(B)}^{2(f_1 v)} - R_{w(A)}^{2(f_1 v)} \\
&= \Delta R_w^{2(f_1 v)}
\end{aligned}
\tag{18}
$$

But LaHuis et al. (2014, p. 443, 447) incorrectly characterized this difference measure as if it were $\Delta R_w^{2(f_1)}$, failing to recognize (as did Hox, 2010, p. 77) that it counts

---

[16]Similarly, as noted in Appendix A, the Raudenbush and Bryk (2002, p. 79) within-cluster measure computed for a single model (compared to a random-intercept-only null) is equivalent in the population to $R_w^{2(f_1 v)}$ (Rights & Sterba [2019]).

measures during model building affords more specific information about the importance of individual terms in the model by providing information on effect size associated with each added term. In this example, wherein 2000 students were nested within 100 classrooms, there was substantive interest in how much variance in student popularity is explained by student sex and extraversion and by teacher years of experience, as well by the cross-level interaction of student extraversion and teacher experience. Although in model building, we could add a single term at a time to the MLM starting with the simplest model, for pedagogical purposes, in two instances we add multiple terms of different kinds simultaneously (from Model A to B and from B to C) in order to illustrate how our measures can nonetheless still identify these terms' unique contributions to explained variance.

We consider six nested models, denoted A, B, C, D, E, and F. Here we implement a *hierarchical* model-building approach, meaning terms are added sequentially, going from the simplest model to the most complex model; this can be contrasted with a *simultaneous* approach described later in the *Extensions* section. In accordance with Step 1 of Table 2 we first clarify what terms are added to each model. Model A is a random-intercept-only model. Model B adds to Model A a fixed component for both the within-class and between-class parts of sex, namely the slope of the level-1 predictor class-mean-centered sex and the level-2 predictor class-mean sex. Model C similarly adds to Model B a fixed component of class-mean-centered extraversion and class-mean extraversion. Model D adds to Model C a fixed slope of level-2 predictor teacher years of experience. Model E adds to Model D a random component of the slope of class-mean-centered extraversion (as well as a covariance between the random intercept and random slope of class-mean-centered extraversion). Lastly, Model F adds a fixed component for the cross-level interaction of class-mean-centered extraversion and teacher experience. In Appendix B, we provide level-specific scalar equation expressions for these Models A-F. All models were fit with restricted maximum likelihood (REML) using SAS Proc MIXED. In accordance with Step 2 of Table 2, we next clarify that we are interested in quantifying the impact of added terms both overall and at a particular level; hence for each model comparison we are interested in both total and level-specific ΔR^2 measures. In accordance with Step 3 of Table 2, we identify the target single-source ΔR^2 measures that reflect the impact of added terms in each model comparison. Results for these target measures are given in Table 3 (where the hats over each ΔR^2 indicate that these are sample estimates). In accordance with Step 4 of Table 2, we then juxtapose and visualize results from all single source measures, in Figure 2.

Specific interpretations of target measure results are as follows. The ΔR^2 between Models A and B helps ascertain the contribution of both the within-class and between-class fixed component of sex (f_1 and f_2, respectively) above and beyond that of the random-intercept-only model. In Table 3 we see that, of the total variance in popularity, an additional 19% is explained by class-mean-centered sex via its fixed component ($\Delta \hat{R}_t^{2(f_1)} = .19$) and an additional 17% is explained by class-mean sex via its fixed component ($\Delta \hat{R}_t^{2(f_2)} = .17$). Further, of the within-cluster variance in popularity, an additional 31% is explained by class-mean-centered sex via its fixed effect ($\Delta \hat{R}_w^{2(f_1)} = .31$) and an additional 45% is explained by class-mean sex via its fixed effect ($\Delta \hat{R}_b^{2(f_2)} = .45$). Thus, there is evidence that sex accounts for a sizable portion of both within- and between-class differences in popularity, with females being more popular within individual classrooms and classrooms with more females having more popular students on average. This is visualized in Figure 2 by comparing Columns 1 and 2.[17]

Going from Model B to Model C helps ascertain the contribution of these same two components (f_1 and f_2) with regards to extraversion. Table 3 shows that, of the total variance, a sizeable additional 12% is explained by class-mean-centered extraversion via its fixed component but a negligible amount (<1%) is explained by class-mean extraversion via its fixed component. Further, Table 3 shows that, of the within-cluster variance, an additional 19% is explained by class-mean-centered extraversion via its fixed component (with more extraverted students being more popular than their less extraverted classmates) but very little (<1%) is explained by class-mean extraversion via its fixed component. This can be visualized in Figure 2 by comparing Columns 2 versus 3.

Next we assess the importance of the level-2 predictor teacher experience. Comparing Models C and D reveals that an additional 7% of total variance is explained by teacher experience ($\Delta \hat{R}_t^{2(f_2)} = .07$), as

---

[17]Note that had we instead entered class-mean-centered sex and class-mean sex individually in separate model comparisons, the target single-source ΔR^2 estimates would be identical to three decimal places. Note also, however, that solely adding a cluster-mean-centered predictor will lead to a small increase in the estimated level-2 random intercept variance (Hoffman, 2015; Snijders & Bosker, 1994) and would thus lead to a slight increase in estimates of $R_t^{2(m)}$. Here we follow recommendations (Hoffman, 2015) to add both the within-cluster and between-cluster portions of level-1 variables simultaneously.

Table 3. ΔR^2 results from the illustrative example model comparison of six multilevel Models A-F implementing the Table 2 procedures (using a hierarchical model-building approach).

	Target effect size measures that reflect the impact of added terms							
	Total measures				Level-specific measures			
	Increment in total variance explained by level-1 predictor via fixed effect	Increment in total variance explained by level-1 predictor via random effect	Increment in total variance explained by level-2 predictor via fixed effect	Increment in total variance explained by cross-level interaction via fixed effect	Increment in level-1 variance explained by level-1 predictor via fixed effect	Increment in level-1 variance explained by level-1 predictor via random effect	Increment in level-2 variance explained by level-2 predictor via fixed effect	Increment in level-1 variance explained by cross-level interaction via fixed effect
Model A: Random intercept only								
Model A vs B: Added fixed components of the slope of level-1 predictor class-mean-centered sex and the slope of level-2 predictor class-mean sex	$\Delta\hat{R}_t^{2(f_1)} = .19$		$\Delta\hat{R}_t^{2(f_2)} = .17$		$\Delta\hat{R}_w^{2(f_1)} = .31$		$\Delta\hat{R}_b^{2(f_2)} = .45$	
Model B vs C: Added fixed components of the slope of level-1 predictor class-mean-centered extraversion and the slope of level-2 predictor class-mean extraversion	$\Delta\hat{R}_t^{2(f_1)} = .12$		$\Delta\hat{R}_t^{2(f_2)} < .01$		$\Delta\hat{R}_w^{2(f_1)} = .19$		$\Delta\hat{R}_b^{2(f_2)} < .01$	
Model C vs D: Added fixed slope for level-2 predictor teacher experience			$\Delta\hat{R}_t^{2(f_2)} = .07$				$\Delta\hat{R}_b^{2(f_2)} = .19$	
Model D vs. E: Added random component to the slope for level-1 predictor class-mean-centered extraversion		$\Delta\hat{R}_t^{2(v)} = .02$				$\Delta\hat{R}_w^{2(v)} = .03$		
Model E vs. F: Added cross-level interaction of class-mean-centered extraversion and teacher experience				$\Delta\hat{R}_t^{2(f_1)} = .02$				$\Delta\hat{R}_w^{2(f_1)} = .03$

Table 4. Parameter estimates and standard errors from illustrative example Model F.

Fixed effects	Est	SE	t
Intercept	1.36	0.59	2.30*
slope of class-mean-centered sex	1.23	0.84	2.22*
slope of class-mean-centered extraversion	0.45	0.02	25.62*
slope of class-mean sex	2.17	0.30	7.31*
slope of class-mean extraversion	0.50	0.12	4.20*
slope of teacher experience	0.09	0.01	6.86*
slope of class-mean-centered extraversion × teacher experience	−0.03	<0.01	−9.54*
Variance components	Est	SE	$z^{\dagger}$
variance of level-2 random intercept residuals	0.27	0.04	6.27*
variance of level-2 random slope residuals for class-mean-centered extraversion	−0.01	0.01	−0.17
covariance of level-2 random intercept residuals w/ level-2 random slope residuals of class-mean-centered extraversion	0.01	0.01	1.27
variance of level-1 residuals	0.55	0.02	29.97*

Notes: Results obtained from Proc MIXED in SAS. *significant, $p < .05$.
[†]Conventional z-tests of variance components are conservative; thus here we employed the alpha-correction approach of Fitzmaurice et al. (2011, p. 209). For a discussion of other alternatives, see Rights and Sterba (2016).

is 19% of between-cluster variance ($\Delta\hat{R}_b^{2(f_2)} = .19$). This can be seen in Figure 2 by comparing Columns 3 and 4.

By comparing Models D and E we can assess the contribution of the random component of class-mean-centered extraversion; it accounts for 2% of total variance ($\Delta\hat{R}_t^{2(v)} = .02$) as well as 3% of within-cluster variance ($\Delta\hat{R}_w^{2(v)} = .03$). This can be visualized in Figure 2 by comparing Columns 4 and 5. Lastly, we assess the importance of the cross-level interaction. Going from Model E to F, this interaction accounts for an additional 2% of total variance ($\Delta\hat{R}_t^{2(f_1)} = .02$) and 3% of within-cluster variance ($\Delta\hat{R}_w^{2(f_1)} = .03$). Inspecting the Figure 2 barchart indicates that across-class slope variability in Model E is instead accounted for by the cross-level interaction in Model F, such that, going from Model E to F, the increase in variance attributable to fixed effects (the target source) is accompanied by an equivalent decrease in variance attributable to random slope variance.[18]

In accordance with Step 5 of Table 2 we could optionally decide to also report some combination-source measures mentioned in Table 1, Columns 4 and 5. For instance, if we wanted to quantify the cumulative impact of sex (above and beyond the random-intercept-only model) via fixed components at both the within-classroom and between-classroom level, we could compute the total combination-source measure $\Delta\hat{R}_t^{2(f_1)} + \Delta\hat{R}_t^{2(f_2)} = \Delta\hat{R}_t^{2(f)}$, which is .36 when comparing Models A to B.

Taken together, the suite of ΔR^2 measures depicted in Figure 2 provides information on practical significance that supplements the traditional presentation

of results. The traditional presentation of results, shown here in Table 4, involves reporting point estimates and standard errors for the parameters of the final Model F. We can see from Table 4 that all of the fixed effects are statistically significant. However, as an example of the utility of reporting ΔR^2 effect sizes in MLM, our measures $\Delta\hat{R}_t^{2(f_1)}$ and $\Delta\hat{R}_t^{2(f_2)}$ provided the additional insight that the estimated contribution of level-1 predictor extraversion via its within-cluster fixed component is more practically significant than that of its between-cluster fixed component.

Discussion

This paper's purpose was to resolve areas of confusion surrounding how ΔR^2 measures can be used as effect size indices in multilevel model comparisons. We identified three limitations of prior simulation studies on ΔR^2 in MLM that had led to misleading or incomplete recommendations for practice. To remedy these limitations and misconceptions, we defined a general set of ΔR^2 measures and then provided a concrete, step-by-step procedure for identifying which measure is relevant to which model comparison, and how that measure can be interpreted in practice. We supplied simulated and analytic demonstrations of the limitations in previous studies on ΔR^2 in MLM and showed how the application of our step-by-step procedures and general set of measures overcomes them. Additionally, we provided and illustrated graphical tools and software that allow researchers to automatically compute and visualize the framework of ΔR^2 measures as an integrated set. Next, we provide recommendations, extensions, and avenues for future research.

---

[18]If from Model E to F we had instead added a cross-level interaction involving a variable with a *fixed* slope, the increase in variance attributable to fixed effects (the target source) would instead be accompanied by a decrease in variance attributable to level-1 residuals (Hoffman, 2015).

Recommendations

Overall summary of recommended practice

To summarize, for a given model comparison, we recommend that researchers implement the procedures in Table 2 to determine which target single-source ΔR^2(s) are able to detect the type of term(s) added to Model A to form Model B. We suggest focusing interpretation on these target single-source ΔR^2 measure(s) as quantitative effect size(s) associated with the addition of particular term(s) to the MLM. The target ΔR^2(s) should also be considered and visualized in the context of the other ΔR^2s in the framework (e.g., using the associated bar chart, as in Figure 2).

Comments on using combination-source versus single-source ΔR^2 for MLMs

Earlier, we explained why reporting a combination-source ΔR^2 in isolation can be misleading and why it is both necessary and sufficient to instead report its constituent single-source ΔR^2s. Nonetheless, when multiple sources of explained variance are of interest in a given application, researchers may at times be motivated to additionally report combination-source ΔR^2s to get an omnibus a summary of the overall impact of added term(s) on all these sources taken together. Additionally, researchers may be motivated to compute combination-source ΔR^2 estimates in order to relate them to previously reported estimates of the same measures from earlier studies, given that preexisting studies have focused moreso on combination-source than single-source ΔR^2. Hence, an overall guideline regarding reporting combination-source ΔR^2s is to always interpret combination-source measures in the context of their single-source constituent measures, and to only report combination-source measures that are consistent with the recommendations in Columns 4 and 5 of Table 1. Violating the guidelines in Columns 4 and 5 of Table 1—such as by reporting $\Delta R_t^{2(fvm)}$ when adding a level-2 predictor in the illustrative example Model C to D comparison or reporting $\Delta R_t^{2(fvm)}$, $\Delta R_t^{2(fv)}$, or $\Delta R_w^{2(f_1v)}$ when adding a cross-level interaction in the illustrative example Model E to F comparison—could be misleading in yielding estimates of 0 even though the level-2 predictor did explain 7% of total outcome variance via its fixed slope and even though the cross-level interaction did explain 2% of total outcome variance via its fixed slope. In the latter inappropriate uses of combination-source measures, increases in the target source of explained variance cancel with decreases in other sources of explained variance. Note that these are examples of the issue detailed in the Limitation 2 section.

Comments on using "conditional" versus "marginal" ΔR^2 for MLMs

Previous MLM literature has drawn a distinction between "conditional" R^2 measures—that consider variance attributable to source "f," along with sources "v" and/or "m," to be explained—and "marginal" measures—that consider only variance attributable to source "f" (i.e., predictors via fixed effects) to be explained (e.g., Edwards et al., 2008; Orelien & Edwards, 2008; Rights & Sterba, 2019; Vonesh & Chinchilli, 1997; Wang & Schaalje, 2009; Xu, 2003). Certain measures in our Table 1 framework can be termed conditional (e.g., $\Delta R_t^{2(fvm)}$) whereas others can be termed marginal (e.g., $\Delta R_t^{2(f)}$). Psychologists tend to be less familiar with the conditional perspective and thus may not recognize that actually they are already using certain conditional measures in practice, for instance, when comparing competing MLMs using Raudenbush and Bryk's [2002, p. 79] measure (which we showed in the Limitation 3 section to be equivalent in the population to $\Delta R_w^{2(f_1v)}$). Conditional measures have been employed across disciplines to assess how accurate a model's predictions are when taking into account *all* cluster-specific information provided by a model—that is, not just information from cluster-level predictors but also from cluster-specific intercepts and/or slopes. In essence, from a conditional perspective, cluster membership is itself an inherent source of explanation when computing R^2. The conditional perspective can be useful, according to Vonesh and Chinchilli (1997, p. 423), because a "moderately low value for [a marginal measure] may mislead the user into thinking the selected fixed effects fit the data poorly. Therefore, it is important that we also assess the fit of both the fixed and random effects based on the conditional mean response" (see also Rights & Sterba, 2019, p. 320).

A benefit of the framework provided in the current paper is that exclusively adopting a marginal or a conditional approach is not necessary. We simply provide an informative partitioning of variance, and show visually how this partitioning changes when adding terms to a model. A researcher can visualize and report changes in this partitioning whether they deem the individual sources of variance to be "explained" or "unexplained." For instance, in the illustrative example, we added a random slope to Model D and computed the target measure $\Delta R_t^{2(v)}$. From a conditional perspective, $\Delta R_t^{2(v)}$ tells us that the random

slope *explains* 2% of the total outcome variance. From a marginal perspective, this measure instead informs us how much variance there is *to be explained* by cross-level interaction terms, such as that introduced in Model F. Lastly, from a neutral perspective, $\Delta R_t^{2(v)}$ is an effect size quantifying the degree of random slope variability and simply tells us that random slope variation accounts for an estimated 2% of the total outcome variance.

Comments on using total versus level-specific ΔR^2 for MLMs

Step 2 of our decision-making framework in Table 2 asked researchers to specify whether their interest focused on total and/or level-specific ΔR^2s. Presently, level-specific measures are more commonly reported in psychology and education applications whereas total measures are more common in biomedical applications. Rather than reflecting distinct disciplinary needs, such reporting differences are likely to simply reflect a lack of awareness across disciplines of the full set of possibilities in Table 1.

Regarding when researchers would find total versus level-specific measures most informative, consider the following examples. Researchers studying math achievement among students nested within schools may find total measures more informative when they wish to explain differences in math achievement across *all* students, that is, both students within the same school and students coming from different schools. Additionally, certain kinds of combination-source measures are only possible as total measures—those that combine sources from different levels (e.g. $\Delta R_t^{2(f)} = \Delta R_t^{2(f_1)} + \Delta R_t^{2(f_2)}$). On the other hand, researchers who want to primarily explain why students from the same school perform differently on math tests may be most interested in within-cluster measures quantifying within-school variance explained. Researchers who instead primarily want to explain why schools have different average levels of performance may be most interested in between-cluster measures quantifying between-school variance explained. However, we argue that inspecting total *and* level-specific measures in juxtaposition for a target source of explained variance in Table 2 affords the most comprehensive context for understanding and interpreting results. For instance, doing so allows identification of situations wherein a small portion of total variance is explained by a given source, despite a lot of level-specific variance being explained by that source.

Extensions

Extension: Using ΔR^2 with simultaneous versus hierarchical MLM model building strategies

In the earlier illustrative example, we computed ΔR^2 for model comparisons using a *hierarchical* model building approach, meaning that terms were added sequentially to build from the most parsimonious MLM (Model A) to the full (most complex) MLM, Model F. A characteristic of a hierarchical approach is that ΔR^2 results depend on the order in which terms are added to the MLM—meaning that, if terms were added in a different order, then the ΔR^2 result for a given term could differ because that ΔR^2 result controls for previously added but not subsequently added terms. The same characteristic holds of the hierarchical approach implemented for single-level regression rather than MLM. Hoffman (2015), Hox (2010), and Snijders and Bosker (2012) provide rationales for particular orderings of terms in hierarchical model building in MLMs.

If the order-dependence characteristic of the hierarchical approach is not desirable in a particular substantive context, an alternative is a *simultaneous* model building approach wherein the full (most complex) MLM is compared to each possible reduced MLM that removes (usually) one term at a time from the full model. Under a simultaneous approach, each MLM ΔR^2 controls for all other terms. Note, however, that the ΔR^2 may not have a substantively useful interpretation for every term under a simultaneous approach (e.g., if the full MLM includes an interaction, a researcher may not want to interpret the ΔR^2 representing the unique contribution of a conditional main effect). The same characteristic holds of the simultaneous approach when implemented in single-level regression rather than MLM (e.g. Cohen, Cohen, West & Aiken, 2003).

To employ the simultaneous approach in the context of the illustrative example MLM, we can designate Model F as the full model, Model G as the full model minus the fixed components for class-mean-centered sex and class-mean sex, Model H as the full model minus the fixed components for class-mean-centered extraversion and class-mean extraversion, Model I as the full model minus the random component for class-mean-centered extraversion, Model J as the full model minus the fixed component for teacher experience, and Model K as the full model minus the fixed component for the cross-level interaction. See Appendix B for scalar level-specific expressions for

Table 5. Extension: Decomposing outcome variance into unique and common predictor-specific contributions by a given source (f_1, f_2, or v) using a simultaneous* model-building approach for the illustrative example.

Source	Predictor	Proportion of total variance	Proportion of within-cluster variance	Proportion of between-cluster variance
f_1	class-mean-centered sex	$\Delta\hat{R}^{2(f_1)}_{t(FG)}=.16$	$\Delta\hat{R}^{2(f_1)}_{w(FG)}=.27$	–
	class-mean-centered extraversion[†]	$\Delta\hat{R}^{2(f_1)}_{t(FH)}=.13$	$\Delta\hat{R}^{2(f_1)}_{w(FH)}=.21$	–
	class-mean-centered extraversion × teacher experience	$\Delta\hat{R}^{2(f_1)}_{t(FK)}=.02$	$\Delta\hat{R}^{2(f_1)}_{w(FK)}=.03$	–
	common	$\hat{R}^{2(f_1)}_{t(F)}-(\Delta\hat{R}^{2(f_1)}_{t(FG)}+\Delta\hat{R}^{2(f_1)}_{t(FH)}+\Delta\hat{R}^{2(f_1)}_{t(FK)})=.02$	$\hat{R}^{2(f_1)}_{w(F)}-(\Delta\hat{R}^{2(f_1)}_{w(FG)}+\Delta\hat{R}^{2(f_1)}_{w(FH)}+\Delta\hat{R}^{2(f_1)}_{w(FK)})=.02$	–
f_2	class-mean sex	$\Delta\hat{R}^{2(f_2)}_{t(FG)}=.08$	–	$\Delta\hat{R}^{2(f_2)}_{b(FG)}=.20$
	class-mean extraversion	$\Delta\hat{R}^{2(f_2)}_{t(FH)}=.02$	–	$\Delta\hat{R}^{2(f_2)}_{b(FH)}=.07$
	teacher experience[†]	$\Delta\hat{R}^{2(f_2)}_{t(FI)}=.07$	–	$\Delta\hat{R}^{2(f_2)}_{b(FI)}=.19$
	common	$\hat{R}^{2(f_2)}_{t(F)}-(\Delta\hat{R}^{2(f_2)}_{t(FG)}+\Delta\hat{R}^{2(f_2)}_{t(FH)}+\Delta\hat{R}^{2(f_2)}_{t(FI)})=.07$	–	$\hat{R}^{2(f_2)}_{b(F)}-(\Delta\hat{R}^{2(f_2)}_{b(FG)}+\Delta\hat{R}^{2(f_2)}_{b(FH)}+\Delta\hat{R}^{2(f_2)}_{b(FI)})=.18$
v	class-mean-centered extraversion	$\Delta\hat{R}^{2(v)}_{t(FI)}<.01$	$\Delta\hat{R}^{2(v)}_{w(FI)}<.01$	–
m	–	$\hat{R}^{2(m)}_{t(F)}=.14$	–	$\hat{R}^{2(m)}_{b(F)}=.36$
level-1 residuals	–	$1-\hat{R}^{2(f_1vm)}_{t(F)}=.29$	$1-\hat{R}^{2(f_1v)}_{w(F)}=.47$	–

Note. The "common" rows provide the estimates of the proportion of variance attributable to the set of predictors jointly by a given source; the rows with specific predictors listed provide estimates of the proportion of variance attributable uniquely to each predictor by a given source. For each ΔR^2 the subscript parenthetical denotes the two models being compared (e.g., Model F vs. Model G for measures in the first row) and for each R^2 the subscript parenthetical denotes the single model for which it is computed (e.g., the full model, Model F, for measures in the last row).
[†]Under a simultaneous approach, the ΔR^2 for every term may not have a substantively useful interpretation (for instance if the full MLM includes an interaction, a researcher may not want to interpret the ΔR^2 representing the unique contribution of a conditional main effect).
*For distinctions between implementing ΔR^2 with a simultaneous (Table 5) versus hierarchical (Table 3) model-building approach, see the discussion section entitled "Extension: Using ΔR^2 with simultaneous versus hierarchical MLM model building strategies."

Models F-K. We can then compute ΔR^2s when comparing Models G vs. F, H vs. F, I vs. F, J vs. F, and K vs. F. Importantly, we can still use the Table 2 procedures to do so, for each of the latter five model comparisons. That is, the Table 2 procedures apply to each pairwise model comparison embedded either within a hierarchical or simultaneous model-building approach.

Extension: If research interest lies in computing "pseudo-R^2" (proportion reduction in residual variance) measures for any null model, there are benefits to using our ΔR^2 framework to do so

Earlier we noted that the Raudenbush & Bryk (2002, p. 79, 74, 85) proportion reduction measures have been called *pseudo-R^2s*. A pseudo-R^2 can be expressed generally as (Hoffman, 2015):

$$pseudo\text{-}R^2 = \frac{\text{residual variance of reduced model-residual variance of full model}}{\text{residual variance of reduced model}} \tag{19}$$

The simultaneous approach also conveniently allows us to decompose outcome variance from the fullest model into that attributable uniquely to specific predictors versus that attributable to the set of predictors jointly, as shown for the illustrative example in Table 5. In single-level regression contexts, the former is often termed the *unique* contribution of the predictors and the latter the *common* contribution of predictors (e.g., Ray-Mukherjee, Nimon, Mukherjee, Morris, Slotow, & Hamer, 2014). In a multilevel context, we can distinguish the unique versus common contribution of predictors for the total/level-specific variance attributable to each of f_1, f_2, and v. For instance, the unique contribution of the fixed component of the level-1 predictor class-mean-centered sex to the full Model F's $R_{t(F)}^{2(f_1)}$ can be computed as $\Delta\hat{R}_{t(FG)}^{2(f_1)}$, wherein the *(FG)* subscript notation, also used in Table 5, refers to subtracting the $R_t^{2(f_1)}$ from Model G from the $R_t^{2(f_1)}$ for Model F. The common contribution of the fixed component of the level-1 predictors is then simply $R_{t(F)}^{2(f_1)}$ for the full Model F minus the sum of the $\Delta R_t^{2(f_1)}$s for each individual term (notated in Table 5 as $R_{t(F)}^{2(f_1)} - (\Delta R_{t(FG)}^{2(f_1)} + \Delta R_{t(FH)}^{2(f_1)} + \Delta R_{t(FK)}^{2(f_1)})$). As shown in Table 5, this same procedure can be done for $R_t^{2(f_2)}$ (computing the unique contribution for each level-2 predictor's fixed slope) and $R_t^{2(v)}$ (computing the unique contribution for each level-1 predictor's random component), and for the level-specific $R_w^{2(f_1)}$, $R_b^{2(f_2)}$, and $R_w^{2(v)}$ (using the level-specific ΔR^2s). In Table 5 the common plus unique contributions of each term are shown to sum down each column to 100% of the total variance (2nd column), 100% of within-cluster variance (3rd column), and 100% of between-cluster variance (4th column) for Model F using the simultaneous approach.

where the residual variance in Equation (19) is either the level-1 residual variance, the level-2 random intercept variance, or a level-2 random slope's variance, depending on the pseudo-R^2 measure. Though we earlier discussed how our framework subsumes two of these three pseudo-R^2 measures from Raudenbush and Bryk (2002, p. 79 and p. 74) when the reduced model is a random-intercept only null model, we now discuss how our framework relates more generally to all three pseudo-R^2s (Raudenbush & Bryk, 2002, p. 79, p, 74, and p. 85) for *any* user-defined null model.

Pseudo-R^2 measures can be computed using our ΔR^2 framework, and moreover there are distinct benefits to doing so, in that our framework allows researchers to simultaneously consider not only (a) the proportion of the reduced model *residual* variance explained by the added term(s), but also (b) the proportion of the *total outcome* variance explained and (c) the proportion of *level-specific outcome* variance explained. Relying solely on (a), as is traditionally done with the pseudo-R^2 measures, can be misleading when there is little residual variance in the reduced model because then a large pseudo-R^2 can be obtained even if the added terms account for little actual outcome variance.

We next show that the pseudo-R^2 measures are simply scaled versions of certain ΔR^2 from our framework. The proportion reduction in level-1 residual variance for any reduced model is defined as

$$pseudo\text{-}R_{L1}^2 = \frac{\sigma_{reduced}^2 - \sigma^2}{\sigma_{reduced}^2} \tag{20}$$

Here $\sigma_{reduced}^2$ is the level-1 residual variance for the reduced model and σ^2 is that for the full model. Appendix C provides a derivation showing that $pseudo\text{-}R_{L1}^2$ is equivalent in the population to scaled versions of our $\Delta R_t^{2(fvm)}$ and $\Delta R_w^{2(f_1v)}$, as described in Equation (21):

$$pseudo\text{-}R_{L1}^2 = \frac{\Delta R_t^{2(fvm)}}{1 - R_{t,\,reduced}^{2(fvm)}}$$

$$= \frac{\Delta R_w^{2(f_1 v)}}{1 - R_{w,\,reduced}^{2(f_1 v)}} \quad (21)$$

Thus our framework allows juxtaposing the *pseudo-R_{L1}^2* (proportion of the reduced model *level-1 residual variance* accounted for by the added terms) with $\Delta R_t^{2(fvm)}$ (proportion of the *total outcome variance* accounted for by the added terms) and with $\Delta R_w^{2(f_1 v)}$ (proportion of the *within-cluster outcome variance* accounted for by the added terms).

Similarly, the proportion reduction in level-2 random intercept variance for any reduced model is:

$$pseudo\text{-}R_{int}^2 = \frac{\tau_{00,\,reduced} - \tau_{00}}{\tau_{00,\,reduced}} \quad (22)$$

$\tau_{00,\,reduced}$ is the level-2 random intercept variance for the reduced model and τ_{00} is that for the full model. In Appendix C we derive the equivalence in the population of *pseudo-R_{int}^2* to scaled versions of $\Delta R_t^{2(m)}$ and $\Delta R_b^{2(m)}$ as shown in Equation (23):

$$pseudo\text{-}R_{int}^2 = \frac{\Delta R_t^{2(m)}}{- R_{t,\,reduced}^{2(m)}}$$

$$= \frac{\Delta R_b^{2(m)}}{- R_{b,\,reduced}^{2(m)}} \quad (23)$$

Using our framework, the *pseudo-R_{int}^2* can thus be juxtaposed with $\Delta R_t^{2(m)}$ and $\Delta R_b^{2(m)}$.

Lastly, the proportion reduction in level-2 random slope variance for any reduced model is:

$$pseudo\text{-}R_{slope}^2 = \frac{\tau_{11,\,reduced} - \tau_{11}}{\tau_{11,\,reduced}} \quad (24)$$

$\tau_{11,\,reduced}$ is a level-2 random slope variance for the reduced model and τ_{11} is that for the full model. In Appendix C, we derive the following equivalence in the population of *pseudo-R_{slope}^2* to scaled versions of $\Delta R_t^{2(v)}$ and $\Delta R_w^{2(v)}$ when there is one random slope in Models A and B:

$$pseudo\text{-}R_{slope}^2 = \frac{\Delta R_t^{2(v)}}{- R_{t,\,reduced}^{2(v)}}$$

$$= \frac{\Delta R_w^{2(v)}}{- R_{w,\,reduced}^{2(v)}} \quad (25)$$

pseudo-R_{slope}^2 is typically reported in isolation after adding a cross-level interaction to the reduced model. However, it is useful to juxtapose it with $\Delta R_t^{2(v)}$ and $\Delta R_w^{2(v)}$ from our framework for reasons we now illustrate. Consider the Model E to F comparison from our earlier illustrative example, in which we added a cross-level interaction term. In this case, the estimate of *pseudo-R_{slope}^2* is 1, suggesting a large effect size associated with the cross-level interaction. Although the addition of the cross-level interaction to Model E indeed accounts for all outcome variance due to random slopes, our framework indicates that the interaction term's importance must be gauged against the fact that there was little outcome variance available to be explained. That is, the random slope accounted for very little outcome variance in the first place ($\Delta \hat{R}_t^{2(v)} = .02$ and $\Delta \hat{R}_w^{2(v)} = .03$ from Model D to E).

Additional extensions and future directions

Next we consider three additional extensions and future directions. First, though we recommend researchers utilize cluster-mean-centering to both disaggregate within-cluster and between-cluster effects and to facilitate decomposition of level-specific proportions of outcome variance, a subset of the Table 1 measures could nonetheless be computed without cluster-mean-centering, i.e., when level-1 variables have both within-cluster and between-cluster variability. In this case, $R_t^{2(f)}$, $R_t^{2(v)}$, $R_t^{2(m)}$, $R_t^{2(fv)}$, and $R_t^{2(fvm)}$ can each be computed for Models A and B using modified formulae provided in Rights and Sterba (2019). Taking these differences then yields non-cluster-mean-centered versions of $\Delta R_t^{2(f)}$, $\Delta R_t^{2(v)}$, $\Delta R_t^{2(m)}$, $\Delta R_t^{2(fv)}$, and $\Delta R_t^{2(fvm)}$. Though they provide less information than our full suite of measures, these versions still allow researchers to separately consider total outcome variance explained via *f, v,* and *m,* rather than relying on preexisting measures that implicitly combine these sources.

Second, in the current paper we presented measures based on MLMs with the most widely used outcome distribution, that is, normal outcomes. However, these measures can be extended for use with generalized linear mixed models (GLMMs) in order to accommodate different outcome types, such as binary outcomes. This can be done by adapting the approach of Nakagawa and Schielzeth (2013), Johnson (2014), and Nakagawa et al. (2017) who provide GLMM versions of $R_t^{2(f)}$ and $R_t^{2(fvm)}$. In contrast to the MLM versions, for GLMM the level-1 residual variance changes depending on the error distribution and link function used. For instance, with binary outcomes using a logit link, σ^2 is replaced with $\pi^2/3$. This same replacement can be done for our set of measures, and thus GLMM versions of the differences in Table 1 can be computed.

Lastly, here we provided formulae to compute point estimates of ΔR^2 for MLMs, but a researcher might additionally be interested in computing

confidence intervals of said differences. For confidence interval computation, a straightforward approach would be to use bootstrapping (for an overview of procedures specific to multilevel data, see Goldstein, 2011). For each bootstrap resample, one can compute each of the available ΔR^2, and can thus obtain an empirical sampling distribution of each individually. Note that when computing the full suite of measures for cluster-mean-centered models, each bootstrap resample would require cluster-mean-centering.

Conclusions

The use of effect size measures has been widely recommended to help social scientists move beyond exclusive reliance on statistical significance when fitting and comparing models (e.g., APA, 2008, 2009; Appelbaum et al., 2018; Harlow, Mulaik, & Steiger, 1997; Kelley & Preacher, 2012; Panter & Sterba, 2011). ΔR^2 measures provide effect size differences on a familiar metric that can be a useful supplement to existing inferential and ranking methods for comparing MLMs, such as likelihood ratio tests and information criteria (Edwards et al., 2008). It is our hope that by resolving confusion surrounding the use of ΔR^2 measures for MLM, providing a more general set of ΔR^2 measures for MLM, and developing a clear step-by-step procedure for choosing and interpreting relevant ΔR^2 measures, researchers comparing MLMs will be better equipped to consider the practical importance of included terms.

Article information

References

American Psychological Association. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, 63, 839–851. doi:810.1037/0003-1066X.1063.1039.1839

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, 73(1), 3–25. doi:10.1037/amp0000191

Aguinis, H., & Culpepper, S. A. (2015). An expanded decision-making procedure for examining cross-level interaction effects with multilevel modeling. *Organizational Research Methods*, 18(2), 155–176. doi:10.1177/1094428114563618

Algina, J., & Swaminathan, H. (2011). Centering in two-level nested designs. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 285–312). New York: Taylor and Francis. doi:10.4324/9780203848852.ch15

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis.* Unpublished manuscript, Stanford University, Stanford Evaluation Consortium, School of Education.

Curran, P. J., Lee, T., Howard, A. L., Lane, S., & MacCallum, R. A. (2012). Disaggregating within-person and between-person effects in multilevel and structural equation growth models. In J. R. Harring & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 217–253). Charlotte, NC: Information Age.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., … Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, 79(1), 69–102. doi:10.3102/0034654308325581

Dimova, R. B., Markatou, M., & Talal, A. H. (2011). Information methods for model selection in linear mixed effects models with application to HCV data. *Computational Statistics & Data Analysis*, 55(9), 2677–2697. doi:10.1016/j.csda.2010.10.031

Dunson, D. B. (2008). *Random effect and latent variable model selection.* New York: Springer. doi:10.1007/978-0-387-76721-5

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. doi:10.1002/sim.3429

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look

at an old issue. *Psychological Methods*, *12*(2), 121–138. doi:10.1037/1082-989X.12.2.121

Fan, Y., & Li, R. (2012). Variable selection in linear mixed effects models. *The Annals of Statistics*, *40*(4), 2043–2068. doi:10.1214/12-AOS1028

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis (2nd ed.)*. Hoboken, N: Wiley.

Goldstein, H. (2011). Bootstrapping in multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 163–172). New York, NY: Routledge. doi:10.4324/9780203848852.ch9

Hamaker, E. L., van Hattum, P., Kuiper, R. M., & Hoijtink, H. (2011). Model selection based on information criteria in multilevel modeling. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 231–255). New York, NY: Taylor & Francis. doi:10.4324/9780203848852.ch13

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum. doi:10.4324/9781315629049

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, *24*(5), 623–641. doi:10.1016/S0149-2063(99)80077-4

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. New York: Routledge.

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications*. (2nd ed.). New York: Routledge.

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R^2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, *44*(6), 1086–1105. doi:10.1080/02664763.2016.1193725

Jaeger, B. C., Edwards, L. J., & Gurka, M. J. (2019). An R^2 statistic for covariance model selection in the linear mixed model. *Journal of Applied Statistics*, *46*(1), 164–184. doi:10.1080/02664763.2018.1466869

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution*, *5*, 944–946. doi:10.1111/2041-210X.12225

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

Kramer, M. (2005, April). *R2 statistics for mixed models. Presented at the 17th Annual Kansas State University Conference on Applied Statistics in Agriculture.* doi:10.4148/2475-7772.1142

Kreft, I. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage. doi: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. doi:10.1177/1094428114541701

Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, *28*(2), 135–167. doi:10.1214/12-STS410

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

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*, 20170213. doi:10.1101/095851

Orelien, J. G., & Edwards, L. J. (2008). Fixed-effect variable selection in linear mixed models using R^2 statistics. *Computational Statistics & Data Analysis*, *52*, 1896–1907. doi:10.1016/j.csda.2007.06.006

Panter, A. T., & Sterba, S. K. (2011). *Handbook of ethics in quantitative methodology*. New York, NY: Routledge. doi:10.4324/9780203840023

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85–112. doi:10.1016/j.jsp.2009.09.002

Potthoff, R. F., & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, *51*(3–4), 313–326. doi:10.1093/biomet/51.3-4.313

Pu, W., & Niu, X. F. (2006). Selecting mixed-effects models based on a generalized information criterion. *Journal of Multivariate Analysis*, *97*(3), 733–758. doi:10.1016/j.jmva.2005.05.009

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Ray-Mukherjee, J., Nimon, K., Mukherjee, S., Morris, D. W., Slotow, R., & Hamer, M. (2014). Using commonality analysis in multiple regressions: A tool to decompose regression effects in the face of multicollinearity. *Methods in Ecology and Evolution*, *5*, 320–328. doi:10.1111/2041-210X.12166

Rights, J. D., & Sterba, S. K. (2016). The relationship between multilevel models and non-parametric multilevel mixture models: Discrete approximation of intraclass correlation, random coefficient distributions, and residual heteroscedasticity. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 316–343. doi:10.1111/bmsp.12073

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), 309hx2013;338. doi:10.1037/met0000184

Ryoo, J. H. (2011). Model selection with the linear mixed model for longitudinal data. *Multivariate Behavioral Research*, *46*(4), 598–624. doi:10.1080/00273171.2011.589264

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, *22*(3), 342–363. doi:10.1177/0049124194022003004

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, *50*(4), 1171–1177. doi:10.2307/2533455

Vonesh, E. F., & Chinchilli, V. M. (1997). *Linear and nonlinear models for the analysis of repeated measurements*. New York: Marcel Dekker. doi:10.1201/9781482293272

Vong, C., Bergstrand, M., Nyberg, J., & Karlsson, M. O. (2012). Rapid sample size calculations for a defined likelihood ratio test-based power in mixed-effects models. *The AAPS Journal*, *14*(2), 176–186. doi:10.1208/s12248-012-9327-8

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. doi:10.1080/03610910802645362

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

Appendices

Appendix A: Review of definitions for multilevel model (MLM) R²s in Rights and Sterba's (2019) integrative framework used for evaluating a single model in isolation: Previous authors of measures equivalent in the population are listed in the table notes

	Symbol	Definition	Computation*
Total MLM R^2s			
Single-source measures	$R_t^{2(f_1)}$	Proportion of total outcome variance explained by *level-1 predictors via fixed components of slopes*	$\frac{\text{Equation (7)}}{\text{Equation (6)}}$
	$R_t^{2(f_2)}$	Proportion of total outcome variance explained by *level-2 predictors via fixed components of slopes*	$\frac{\text{Equation (8)}}{\text{Equation (6)}}$
	$R_t^{2(v)}$	Proportion of total outcome variance explained by *level-1 predictors via random slope (co)variation*[1]	$\frac{\text{Equation (9)}}{\text{Equation (6)}}$
	$R_t^{2(m)}$	Proportion of total outcome variance explained by *cluster-specific outcome means via random intercept variation*	$\frac{\text{Equation (10)}}{\text{Equation (6)}}$
Combination-source measures	$R_t^{2(f)}$	Proportion of total outcome variance explained by *all predictors via fixed slopes*[2]	$R_t^{2(f_1)} + R_t^{2(f_2)}$
	$R_t^{2(fv)}$	Proportion of total outcome variance explained by *predictors via fixed slopes & random slope (co)variation*	$R_t^{2(f_1)} + R_t^{2(f_2)} + R_t^{2(v)}$
	$R_t^{2(fvm)}$	Proportion of total outcome variance explained by *predictors via fixed slopes and random slope (co)variation* & by *cluster-specific outcome means via random intercept variation*[3]	$R_t^{2(f_1)} + R_t^{2(f_2)} + R_t^{2(v)} + R_t^{2(m)}$
Within-cluster MLM R^2s			
Single-source measures	$R_w^{2(f_1)}$	Proportion of within-cluster outcome variance explained by *level-1 predictors via fixed components of slopes*	$\frac{\text{Equation (7)}}{\text{Equation (12)}}$
	$R_w^{2(v)}$	Proportion of within-cluster outcome variance explained by *level-1 predictors via random slope (co)variation*	$\frac{\text{Equation (9)}}{\text{Equation (12)}}$
Combination-source measure	$R_w^{2(f_1 v)}$	Proportion of within-cluster outcome variance explained by *level-1 predictors via fixed slopes & random slope (co)variation*[4]	$R_w^{2(f_1)} + R_w^{2(v)}$
Between-cluster MLM R^2s			
Single-source measures	$R_b^{2(f_2)}$	Proportion of between-cluster outcome variance explained by *level-2 predictors via fixed components of slopes*[4]	$\frac{\text{Equation (8)}}{\text{Equation (13)}}$
	$R_b^{2(m)}$	Proportion of between-cluster outcome variance explained by *cluster-specific outcome means via random intercept variation*	$\frac{\text{Equation (10)}}{\text{Equation (13)}}$

Notes. *= Computation refers to equation numbers in the current paper. Authors of a measure equivalent in the population are as follows: 1= Aguinis & Culpepper (2015); 2 = Johnson (2014) (extension of Nakagawa & Schielzeth [2013]) and Snijders & Bosker (2012) and Vonesh & Chinchilli (1997); 3 = Johnson (2014) (extension of Nakagawa & Schielzeth [2013]) and Vonesh & Chinchilli (1997) and Xu (2003); 4 = Hox (2010) and Kreft & de Leeuw (1998) and Vonesh & Chinchilli (1997) and Raudenbush & Bryk's (2002) "pseudo-R^2s" (computed with a random-intercept-only null model). See Rights & Sterba (2019, Appendix B) for derivations underlying each of these population equivalencies.

Appendix B: Scalar model expressions for the illustrative example

In the manuscript section titled *Illustrative Example*, we described Models A-F that were used in a hierarchical model building approach. In the subsection titled *Extension: Using ΔR^2 with simultaneous versus hierarchical MLM model building strategies*, we described Models F-K that were used in a simultaneous model building approach with the same data. Here, we supplement all of the manuscript text descriptions of Models A-K with scalar level-1 and level-2 expressions (as in manuscript Equation (1)–(2)). For each model below, *popular$_{ij}$* denotes student *i*'s popularity (within class *j*); additionally, all level-1 residuals (e_{ij}'s) are normally distributed, and all random effect residuals are multivariate normally distributed (with all random effect residuals allowed to covary).

Model A:

$$popular_{ij} = \beta_{0j} + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Model B:

$$y_{ij} = \beta_{0j} + \beta_{1j}(sex_{ij}-sex_{\cdot j}) + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}sex_{\cdot j} + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$

(where $sex_{ij}-sex_{\cdot j}$ = class-mean-centered student sex and $sex_{\cdot j}$ = class-mean sex)

Model C:
$$y_{ij} = \beta_{0j} + \beta_{1j}(sex_{ij}-sex_{\cdot j}) + \beta_{2j}(extrav_{ij}-extrav_{\cdot j}) + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}sex_{\cdot j} + \gamma_{02}extrav_{\cdot j} + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20}$$

(where $extrav_{ij}-extrav_{\cdot j}$ = class-mean-centered student extraversion and $extrav_{\cdot j}$ = class-mean extraversion)

Model D:
$$y_{ij} = \beta_{0j} + \beta_{1j}(sex_{ij}-sex_{\cdot j}) + \beta_{2j}(extrav_{ij}-extrav_{\cdot j}) + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}sex_{\cdot j} + \gamma_{02}extrav_{\cdot j} + \gamma_{03}teach\,exp_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20}$$

(where $teach\,exp_j$ = teacher years of experience)

Model E:
$$y_{ij} = \beta_{0j} + \beta_{1j}(sex_{ij} - sex_{\cdot j}) + \beta_{2j}(extrav_{ij} - extrav_{\cdot j}) + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}sex_{\cdot j} + \gamma_{02}extrav_{\cdot j} + \gamma_{03}teach\,exp_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20} + u_{2j}$$

Model F:
$$y_{ij} = \beta_{0j} + \beta_{1j}(sex_{ij} - sex_{\cdot j}) + \beta_{2j}(extrav_{ij} - extrav_{\cdot j}) + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}sex_{\cdot j} + \gamma_{02}extrav_{\cdot j} + \gamma_{03}teach\,exp_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}teach\,exp_j + u_{2j}$$

Model G:
$$y_{ij} = \beta_{0j} + \beta_{1j}(extrav_{ij} - extrav_{\cdot j}) + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}extrav_{\cdot j} + \gamma_{02}teach\,exp_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}teach\,exp_j + u_{2j}$$

Model H:
$$y_{ij} = \beta_{0j} + \beta_{1j}(sex_{ij} - sex_{\cdot j}) + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}sex_{\cdot j} + \gamma_{02}teach\,exp_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}teach\,exp_j + u_{2j}$$

Model I:
$$y_{ij} = \beta_{0j} + \beta_{1j}(sex_{ij} - sex_{\cdot j}) + \beta_{2j}(extrav_{ij} - extrav_{\cdot j}) + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}sex_{\cdot j} + \gamma_{02}extrav_{\cdot j} + \gamma_{03}teach\,exp_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}teach\,exp_j$$

Model J:
$$y_{ij} = \beta_{0j} + \beta_{1j}(sex_{ij} - sex_{\cdot j}) + \beta_{2j}(extrav_{ij} - extrav_{\cdot j}) + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}sex_{\cdot j} + \gamma_{02}extrav_{\cdot j} + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}teach\,exp_j + u_{2j}$$

Model K:
$$y_{ij} = \beta_{0j} + \beta_{1j}(sex_{ij} - sex_{\cdot j}) + \beta_{2j}(extrav_{ij} - extrav_{\cdot j}) + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}sex_{\cdot j} + \gamma_{02}extrav_{\cdot j} + \gamma_{03}teach\,exp_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20} + u_{2j}$$

Appendix C

Analytically relating "pseudo-R^2" measures to our ΔR^2 framework

Derivations showing the equivalence in the population of manuscript equations (20) and (21) are provided here in Appendix C equations (C1) and (C2). Symbols were defined in the manuscript text.

$$
\begin{aligned}
pseudo\text{-}R_{L1}^2 &= \frac{\sigma_{reduced}^2 - \sigma^2}{\sigma_{reduced}^2} \\[2mm]
&= \frac{(\sigma_{reduced}^2 - \sigma^2)/(\text{total outcome variance})}{\sigma_{reduced}^2/(\text{total outcome variance})} \\[2mm]
&= \frac{\dfrac{\sigma_{reduced}^2}{\text{total outcome variance}} - \dfrac{\sigma^2}{\text{total outcome variance}}}{\dfrac{\sigma_{reduced}^2}{\text{total outcome variance}}} \\[4mm]
&= \frac{\left(1 - \dfrac{\gamma_{reduced}^w{}'\mathbf{\Phi}^w\gamma_{reduced}^w + \gamma_{reduced}^b{}'\mathbf{\Phi}^b\gamma_{reduced}^b + tr(\mathbf{T\Sigma})_{reduced} + \tau_{00,\,reduced}}{\text{total outcome variance}}\right) - \left(1 - \dfrac{\gamma^{w\prime}\mathbf{\Phi}^w\gamma^w + \gamma^{b\prime}\mathbf{\Phi}^b\gamma^b + tr(\mathbf{T\Sigma}) + \tau_{00}}{\text{total outcome variance}}\right)}{\left(1 - \dfrac{\gamma_{reduced}^w{}'\mathbf{\Phi}^w\gamma_{reduced}^w + \gamma_{reduced}^b{}'\mathbf{\Phi}^b\gamma_{reduced}^b + tr(\mathbf{T\Sigma})_{reduced} + \tau_{00,\,reduced}}{\text{total outcome variance}}\right)} \\[4mm]
&= \frac{\left(1 - R_{t,\,reduced}^{2(fvm)}\right) - \left(1 - R_t^{2(fvm)}\right)}{\left(1 - R_{t,\,reduced}^{2(fvm)}\right)} \\[2mm]
&= \frac{R_t^{2(fvm)} - R_{t,\,reduced}^{2(fvm)}}{\left(1 - R_{t,\,reduced}^{2(fvm)}\right)} \\[2mm]
&= \frac{\Delta R_t^{2(fvm)}}{\left(1 - R_{t,\,reduced}^{2(fvm)}\right)}
\end{aligned}
$$

$$\text{(C1)}$$

$$pseudo\text{-}R^2_{L1} = \frac{\sigma^2_{reduced} - \sigma^2}{\sigma^2_{reduced}}$$

$$= \frac{(\sigma^2_{reduced} - \sigma^2)/(\text{within-cluster outcome variance})}{\sigma^2_{reduced}/(\text{within-cluster outcome variance})}$$

$$= \frac{\dfrac{\sigma^2_{reduced}}{\text{within-cluster outcome variance}} - \dfrac{\sigma^2}{\text{within-cluster outcome variance}}}{\dfrac{\sigma^2_{reduced}}{\text{within-cluster outcome variance}}}$$

$$= \frac{\left(1 - \dfrac{\gamma^w_{reduced}{}' \mathbf{\Phi}^w \gamma^w_{reduced} + tr(\mathbf{T\Sigma})_{reduced}}{\text{within-cluster outcome variance}}\right) - \left(1 - \dfrac{\gamma^{w\prime} \mathbf{\Phi}^w \gamma^w + tr(\mathbf{T\Sigma})}{\text{within-cluster outcome variance}}\right)}{\left(1 - \dfrac{\gamma^w_{reduced}{}^\beta \mathbf{\Phi}^w \gamma^w_{reduced} + tr(\mathbf{T\Sigma})_{reduced}}{\text{within-cluster outcome variance}}\right)} \quad (C2)$$

$$= \frac{\left(1 - R^{2(f_1 v)}_{w, reduced}\right) - \left(1 - R^{2(f_1 v)}_w\right)}{\left(1 - R^{2(f_1 v)}_{w, reduced}\right)}$$

$$= \frac{R^{2(f_1 v)}_w - R^{2(f_1 v)}_{w, reduced}}{\left(1 - R^{2(f_1 v)}_{w, reduced}\right)}$$

$$= \frac{\Delta R^{2(f_1 v)}_w}{\left(1 - R^{2(f_1 v)}_{w, reduced}\right)}$$

Derivations showing the equivalence in the population of manuscript equations (22) and (23) are provided here in Appendix C equations (C3) and (C4). Symbols were defined in the manuscript text.

$$pseudo\text{-}R^2_{int} = \frac{\tau_{00, reduced} - \tau_{00}}{\tau_{00, reduced}}$$

$$= \frac{(\tau_{00, reduced} - \tau_{00})/(\text{total outcome variance})}{\tau_{00, reduced}/(\text{total outcome variance})}$$

$$= \frac{\dfrac{\tau_{00, reduced}}{\text{total outcome variance}} - \dfrac{\tau_{00}}{\text{total outcome variance}}}{\dfrac{\tau_{00, reduced}}{\text{total outcome variance}}}$$

$$= \frac{R^{2(m)}_{t, reduced} - R^{2(m)}_t}{R^{2(m)}_{t, reduced}}$$

$$= \frac{R^{2(m)}_t - R^{2(m)}_{t, reduced}}{-R^{2(m)}_{t, reduced}}$$

$$= \frac{\Delta R^{2(m)}_t}{-R^{2(m)}_{t, reduced}}$$

$$(C3)$$

$$pseudo\text{-}R^2_{int} = \frac{\tau_{00, reduced} - \tau_{00}}{\tau_{00, reduced}}$$

$$= \frac{(\tau_{00, reduced} - \tau_{00})/(\text{between-cluster outcome variance})}{\tau_{00, reduced}/(\text{between-cluster outcome variance})}$$

$$= \frac{\dfrac{\tau_{00, reduced}}{\text{between-cluster outcome variance}} - \dfrac{\tau_{00}}{\text{between-cluster outcome variance}}}{\dfrac{\tau_{00, reduced}}{\text{between-cluster outcome variance}}}$$

$$= \frac{R^{2(m)}_{b, reduced} - R^{2(m)}_b}{R^{2(m)}_{b, reduced}}$$

$$= \frac{R^{2(m)}_b - R^{2(m)}_{b, reduced}}{-R^{2(m)}_{b, reduced}}$$

$$= \frac{\Delta R^{2(m)}_b}{-R^{2(m)}_{b, reduced}}$$

$$(C4)$$

Derivations showing the equivalence in the population of manuscript equations (24) and (25) when there is one random slope in the both the full and reduced models are provided here in Appendix C equations (C5) and (C6). Symbols were defined in the manuscript text.

$$pseudo\text{-}R^2_{slope} = \frac{\tau_{11,\,reduced} - \tau_{11}}{\tau_{11,\,reduced}}$$

$$= \frac{\text{var}(x_{ij})(\tau_{11,\,reduced} - \tau_{11})/(\text{total outcome variance})}{\text{var}(x_{ij})\tau_{11,\,reduced}/(\text{total outcome variance})}$$

$$= \frac{\dfrac{\text{var}(x_{ij})\tau_{11,\,reduced}}{\text{total outcome variance}} - \dfrac{\text{var}(x_{ij})\tau_{11}}{\text{total outcome variance}}}{\dfrac{\text{var}(x_{ij})\tau_{11,\,reduced}}{\text{total outcome variance}}}$$

$$= \frac{R^{2(v)}_{t,\,reduced} - R^{2(v)}_{t}}{R^{2(v)}_{t,\,reduced}}$$

$$= \frac{R^{2(v)}_{t} - R^{2(v)}_{t,\,reduced}}{-R^{2(v)}_{t,\,reduced}}$$

$$= \frac{\Delta R^{2(v)}_{t}}{-R^{2(v)}_{t,\,reduced}}$$

$$(C5)$$

$$pseudo\text{-}R^2_{slope} = \frac{\tau_{11,\,reduced} - \tau_{11}}{\tau_{11,\,reduced}}$$

$$= \frac{\text{var}(x_{ij})(\tau_{11,\,reduced} - \tau_{11})/(\text{within-cluster outcome variance})}{\text{var}(x_{ij})\tau_{11,\,reduced}/(\text{within-cluster outcome variance})}$$

$$= \frac{\dfrac{\text{var}(x_{ij})\tau_{11,\,reduced}}{\text{within-cluster outcome variance}} - \dfrac{\text{var}(x_{ij})\tau_{11}}{\text{within-cluster outcome variance}}}{\dfrac{\text{var}(x_{ij})\tau_{11,\,reduced}}{\text{within-cluster outcome variance}}}$$

$$= \frac{R^{2(v)}_{w,\,reduced} - R^{2(v)}_{w}}{R^{2(v)}_{w,\,reduced}}$$

$$= \frac{R^{2(v)}_{w} - R^{2(v)}_{w,\,reduced}}{-R^{2(v)}_{w,\,reduced}}$$

$$= \frac{\Delta R^{2(v)}_{w}}{-R^{2(v)}_{w,\,reduced}}$$

$$(C6)$$

Online Appendix to accompany "New recommendations on the use of R-squared differences in multilevel model comparisons": *r2MLMcomp* R function

r2MLMcomp **R function Description:**

      This function reads in raw data and multilevel model (MLM) parameter estimates from two separate models under comparison (Model A and Model B) and outputs all R^2 measures (reviewed and defined in the manuscript's *print* Appendix A) for both models as well as all ΔR^2 measures (defined in the manuscript's Table 1). Additionally, the function produces side-by-side graphical comparisons of the R^2 measures for Model A vs. B that can be used to visualize changes in each measure across models (see manuscript description of Figure 2 for an example illustration). This function assumes all level-1 predictors are cluster-mean-centered for reasons described in the manuscript. Any number of level-1 and/or level-2 predictors is supported and any of the level-1 predictors can have random slopes. This function can be used with either the hierarchical or the simultaneous model-building approach described in the manuscript (in which Model B is the more complex model for a given pairwise comparison). Our function can also be used with either nested non-nested model comparisons (in which R^2 estimates for Model A are subtracted from those for Model B).

r2MLMcomp **R function Input:**

data – Dataset with rows denoting observations and columns denoting variables

within_covs_modA – List of numbers corresponding to the columns in the dataset of the level-1 predictors used in the Model A MLM (if none used, set to NULL)

between_covs_modA – List of numbers corresponding to the columns in the dataset of the level-2 predictors used in the Model A MLM (if none used, set to NULL)

random_covs_modA – List of numbers corresponding to the columns in the dataset of the level-1 predictors that have random slopes in the Model A MLM (if no random slopes, set to NULL)

gamma_w_modA – Vector of estimates of the fixed component of slopes for all level-1 predictors for Model A, to be entered in the order of the predictors listed by *within_covs* (if none, set to NULL)

gamma_b_modA – Vector of estimates of the fixed component of the intercept and slopes for all level-2 predictors for Model A, to be entered intercept first followed by level-2 slopes in the order listed by *between_covs*

Tau_modA – random effect covariance matrix for Model A; note that the first row/column denotes the intercept variance and covariances (if intercept is fixed, set all to 0) and each subsequent row/column denotes a given random slope's variance and covariances (to be entered in the order listed by *random_covs*). Variances are on diagonal and covariances are on off-diagonal.

sigma2_modA – level-1 residual variance for Model A

within_covs_modB – List of numbers corresponding to the columns in the dataset of the level-1 predictors used in the Model B MLM (if none used, set to NULL)

between_covs_modB – List of numbers corresponding to the columns in the dataset of the level-2 predictors used in the Model B MLM (if none used, set to NULL)

random_covs_modB – List of numbers corresponding to the columns in the dataset of the level-1 predictors that have random slopes in the Model B MLM (if no random slopes, set to NULL)

gamma_w_modB – Vector of estimates of the fixed component of slopes for all level-1 predictors for Model B, to be entered in the order of the predictors listed by *within_covs* (if none, set to NULL)

gamma_b_modB – Vector of estimates of the fixed component of the intercept and slopes for all

level-2 predictors for Model B, to be entered intercept first followed by level-2 slopes in the order listed by *between_covs*

Tau_modB – random effect covariance matrix for Model B; note that the first row/column denotes the intercept variance and covariances (if intercept is fixed, set all to 0) and each subsequent row/column denotes a given random slope's variance and covariances (to be entered in the order listed by *random_covs*). Variances are on diagonal and covariances are on off-diagonal.

sigma2_modB – level-1 residual variance for Model B

r2MLMcomp R function Code:

```
r2MLMcomp <- function(data,within_covs_modA,between_covs_modA,random_covs_modA,
                gamma_w_modA,gamma_b_modA,Tau_modA,sigma2_modA,
                within_covs_modB,between_covs_modB,random_covs_modB,
                gamma_w_modB,gamma_b_modB,Tau_modB,sigma2_modB){
  ##r2MLM function
  r2MLM <- function(data,within_covs,between_covs,random_covs,
            gamma_w,gamma_b,Tau,sigma2,modelname){
    if(length(gamma_b)>1) gamma <- c(1,gamma_w,gamma_b[2:length(gamma_b)])
   if(length(gamma_b)==1) gamma <- c(1,gamma_w)
   if(is.null(within_covs)==T) gamma_w <- 0
  if(is.null(gamma)) gamma <- 0
  ##compute phi
  phi <- var(cbind(1,data[,c(within_covs)],data[,c(between_covs)]),na.rm=T)
  phi_w <- var(data[,within_covs],na.rm=T)
  if(is.null(within_covs)==T) phi_w <- 0
  phi_b <- var(cbind(1,data[,between_covs]),na.rm=T)
  if(is.null(between_covs)==T) phi_b <- 0
  ##compute psi and kappa
  var_randomcovs <- var(cbind(1,data[,c(random_covs)]),na.rm=T)
    if(length(Tau)>1) psi <- matrix(c(diag(Tau)),ncol=1)
  if(length(Tau)==1) psi <- Tau
    if(length(Tau)>1) kappa <- matrix(c(Tau[lower.tri(Tau)==TRUE]),ncol=1)
  if(length(Tau)==1) kappa <- 0
    v <- matrix(c(diag(var_randomcovs)),ncol=1)
  r <- matrix(c(var_randomcovs[lower.tri(var_randomcovs)==TRUE]),ncol=1)
    if(is.null(random_covs)==TRUE){
   v <- 0
   r <- 0
   m <- matrix(1,ncol=1)
  }
 if(length(random_covs)>0) m <- matrix(c(colMeans(cbind(1,data[,c(random_covs)]),na.rm=T)),ncol=1)
    ##total variance
    totalvar_notdecomp <- t(v)%*%psi + 2*(t(r)%*%kappa) + t(gamma)%*%phi%*%gamma + t(m)%*%Tau%*%m + sigma2
    totalwithinvar <- (t(gamma_w)%*%phi_w%*%gamma_w) + (t(v)%*%psi + 2*(t(r)%*%kappa)) + sigma2
    totalbetweenvar <- (t(gamma_b)%*%phi_b%*%gamma_b) + Tau[1]
    totalvar <- totalwithinvar + totalbetweenvar
    ##total decomp
    decomp_fixed_notdecomp <- (t(gamma)%*%phi%*%gamma) / totalvar
    decomp_fixed_within <- (t(gamma_w)%*%phi_w%*%gamma_w) / totalvar
    decomp_fixed_between <- (t(gamma_b)%*%phi_b%*%gamma_b) / totalvar
    decomp_fixed <- decomp_fixed_within + decomp_fixed_between
    decomp_varslopes <- (t(v)%*%psi + 2*(t(r)%*%kappa)) / totalvar
    decomp_varmeans <- (t(m)%*%Tau%*%m) / totalvar
    decomp_sigma <- sigma2/totalvar
    ##within decomp
    decomp_fixed_within_w <- (t(gamma_w)%*%phi_w%*%gamma_w) / totalwithinvar
    decomp_varslopes_w <- (t(v)%*%psi + 2*(t(r)%*%kappa)) / totalwithinvar
    decomp_sigma_w <- sigma2/totalwithinvar
    ##between decomp
    decomp_fixed_between_b <- (t(gamma_b)%*%phi_b%*%gamma_b) / totalbetweenvar
    decomp_varmeans_b <- Tau[1] / totalbetweenvar
    ##measures
     R2_f <- decomp_fixed
  R2_f1 <- decomp_fixed_within
  R2_f2 <- decomp_fixed_between
```

```
      R2_fv <- decomp_fixed + decomp_varslopes
      R2_fvm <- decomp_fixed + decomp_varslopes + decomp_varmeans
      R2_v <- decomp_varslopes
      R2_m <- decomp_varmeans
      R2_f_w <- decomp_fixed_within_w
      R2_f_b <- decomp_fixed_between_b
      R2_fv_w <- decomp_fixed_within_w + decomp_varslopes_w
      R2_v_w <- decomp_varslopes_w
      R2_m_b <- decomp_varmeans_b
       decomp_table <- matrix(c(decomp_fixed_within,decomp_fixed_between,decomp_varslopes,decomp_varmeans,decomp_sigma,
                  decomp_fixed_within_w,"NA",decomp_varslopes_w,"NA",decomp_sigma_w,
                  "NA",decomp_fixed_between_b,"NA",decomp_varmeans_b,"NA"),ncol=3)
      rownames(decomp_table) <- c("fixed, within","fixed, between","slope variation","mean variation","sigma2")
      colnames(decomp_table) <- c("total","within","between")
      R2_table <- matrix(c(R2_f1,R2_f2,R2_v,R2_m,R2_f,R2_fv,R2_fvm,
                  R2_f_w,"NA",R2_v_w,"NA","NA",R2_fv_w,"NA",
                  "NA",R2_f_b,"NA",R2_m_b,"NA","NA","NA")
                  ,ncol=3)
      rownames(R2_table) <- c("f1","f2","v","m","f","fv","fvm")
      colnames(R2_table) <- c("total","within","between")
    ##barchart
     contributions_stacked <- matrix(c(decomp_fixed_within,decomp_fixed_between,decomp_varslopes,decomp_varmeans,decomp_sigma,
                      decomp_fixed_within_w,0,decomp_varslopes_w,0,decomp_sigma_w,
                      0,decomp_fixed_between_b,0,decomp_varmeans_b,0),5,3)
      colnames(contributions_stacked) <- c("total","within","between")
      rownames(contributions_stacked) <- c("fixed slopes (within)",
                          "fixed slopes (between)",
                          "slope variation (within)",
                          "intercept variation (between)",
                          "residual (within)")
      barplot(contributions_stacked, main=paste0("Decomposition of Scaled Variance, Model ",modelname), horiz=FALSE,
          ylim=c(0,1),col=c("darkred","steelblue","darkred","midnightblue","white"),ylab="proportion of variance",
          density=c(NA,NA,30,40,NA),angle=c(0,45,0,135,0),xlim=c(0,1),width=c(.3,.3))
      legend(.33,-.1,legend=rownames(contributions_stacked),fill=c("darkred","steelblue","darkred","midnightblue","white"),
          cex=.7, pt.cex = 1,xpd=TRUE,density=c(NA,NA,30,40,NA),angle=c(0,45,0,135,0))
     Output <- list(noquote(decomp_table),noquote(R2_table))
     names(Output) <- c("Decompositions","R2s")
     return(Output)
  }
  ##compute decomp for Model A and B
  results_modA <- r2MLM(data,within_covs_modA,between_covs_modA,random_covs_modA,
              gamma_w_modA,gamma_b_modA,Tau_modA,sigma2_modA,"A")
  decomp_modA <- results_modA$Decompositions
  results_modB <- r2MLM(data,within_covs_modB,between_covs_modB,random_covs_modB,
              gamma_w_modB,gamma_b_modB,Tau_modB,sigma2_modB,"B")
  decomp_modB <- results_modB$Decompositions
  ##comparison measures
  delta_f1_t <- as.numeric(decomp_modA[1,1]) - as.numeric(decomp_modB[1,1])
  delta_f2_t <- as.numeric(decomp_modA[2,1]) - as.numeric(decomp_modB[2,1])
  delta_v_t <- as.numeric(decomp_modA[3,1]) - as.numeric(decomp_modB[3,1])
  delta_m_t <-as.numeric(decomp_modA[4,1]) - as.numeric(decomp_modB[4,1])
  delta_f1_w <- as.numeric(decomp_modA[1,2]) - as.numeric(decomp_modB[1,2])
  delta_v_w <- as.numeric(decomp_modA[3,2]) - as.numeric(decomp_modB[3,2])
  delta_f2_b <- as.numeric(decomp_modA[2,3]) - as.numeric(decomp_modB[2,3])
  delta_m_b <- as.numeric(decomp_modA[4,3]) - as.numeric(decomp_modB[4,3])
  delta_f_t <- delta_f1_t + delta_f2_t
  delta_fv_t <- delta_f1_t + delta_f2_t + delta_v_t
  delta_fvm_t <- delta_f1_t + delta_f2_t + delta_v_t + delta_m_t
  delta_f1v_w <- delta_f1_w + delta_v_w
  ##comparison bar charts
 contributions_stacked_total <-
matrix(c(as.numeric(decomp_modA[1,1]),as.numeric(decomp_modA[2,1]),as.numeric(decomp_modA[3,1]),as.numeric(decomp_modA[4,1]),as.
numeric(decomp_modA[5,1]),
as.numeric(decomp_modB[1,1]),as.numeric(decomp_modB[2,1]),as.numeric(decomp_modB[3,1]),as.numeric(decomp_modB[4,1]),as.numeric(d
ecomp_modB[5,1])),5,2)
  colnames(contributions_stacked_total) <- c("Model A","Model B")
   barplot(contributions_stacked_total, main="Decomposition of Scaled Total Variance", horiz=FALSE,
      ylim=c(0,1),col=c("darkred","steelblue","darkred","midnightblue","white"),ylab="proportion of variance",
      density=c(NA,NA,30,40,NA),angle=c(0,45,0,135,0),width=c(.3,.3))
   legend(0.26,-.1,legend=c("fixed slopes (within)",
```

```
                    "fixed slopes (between)",
                    "slope variation (within)",
                    "intercept variation (between)",
                    "residual (within)"),fill=c("darkred","steelblue","darkred","midnightblue","white"),
         cex=.7, pt.cex = 1,xpd=TRUE,density=c(NA,NA,30,40,NA),angle=c(0,45,0,135,0))
    contributions_stacked_within <-
matrix(c(as.numeric(decomp_modA[1,2]),0,as.numeric(decomp_modA[3,2]),0,as.numeric(decomp_modA[5,2]),
                        as.numeric(decomp_modB[1,2]),0,as.numeric(decomp_modB[3,2]),0,as.numeric(decomp_modB[5,2])),5,2)
    colnames(contributions_stacked_within) <- c("Model A","Model B")
    barplot(contributions_stacked_within, main="Decomposition of Scaled Within-Cluster Variance", horiz=FALSE,
         ylim=c(0,1),col=c("darkred","steelblue","darkred","midnightblue","white"),ylab="proportion of variance",
         density=c(NA,NA,30,40,NA),angle=c(0,45,0,135,0),width=c(.3,.3))
    legend(0.28,-.1,legend=c("fixed slopes (within)",
                    "slope variation (within)",
                    "residual (within)"),fill=c("darkred","darkred","white"),
         cex=.7, pt.cex = 1,xpd=TRUE,density=c(NA,30,NA),angle=c(0,0,0))
    contributions_stacked_between <-  matrix(c(0,as.numeric(decomp_modA[2,3]),0,as.numeric(decomp_modA[4,3]),0,
                        0,as.numeric(decomp_modB[2,3]),0,as.numeric(decomp_modB[4,3]),0),5,2)
    colnames(contributions_stacked_between) <- c("Model A","Model B")
    barplot(contributions_stacked_between, main="Decomposition of Scaled Between-Cluster Variance", horiz=FALSE,
         ylim=c(0,1),col=c("darkred","steelblue","darkred","midnightblue","white"),ylab="proportion of variance",
         density=c(NA,NA,30,40,NA),angle=c(0,45,0,135,0),width=c(.3,.3))
    legend(0.26,-.1,legend=c("fixed slopes (between)",
                    "intercept variation (between)"),fill=c("steelblue","midnightblue"),
         cex=.7, pt.cex = 1,xpd=TRUE,density=c(NA,40),angle=c(45,135))
    ##table of R2 deltas
    R2_modA <- results_modA$R2s
    R2_modB <- results_modB$R2s
    R2_delta <- suppressWarnings(as.numeric(R2_modB) - as.numeric(R2_modA))
    R2_delta <- matrix(R2_delta,7,3)
    colnames(R2_delta) <- colnames(R2_modA)
    rownames(R2_delta) <- rownames(R2_modA)
    Output <- list(R2_modA,R2_modB,R2_delta)
    names(Output) <- c("Model A R2s","Model B R2s","R2 differences, Model B - Model A")
      return(Output)
  }
```

r2MLMcomp R function Example Input:

```
#NOTE: estimates in the input represent hypothetical results for a comparison between a random slope model with two
#level-1 predictors and one level-2 predictor (Model A) and a model that adds an additional level-2 predictor (Model B);
#in practice a user would have previously obtained these input estimates by fitting their models in MLM software;
#additionally, the input consists of hypothetical predictor data, whereas in practice a user would read-in their actual data

data <- matrix(NA,100,4)
xs <- mvrnorm(n=100,mu=c(0,0),Sigma=matrix(c(2,.75,.75,1.5),2,2))
ws <- mvrnorm(n=10,mu=c(0,2),Sigma=matrix(c(1,.5,.5,2),2,2))
data[,1:2] <- xs
for (i in seq(10)){
 data[(10*(i-1)+1):(i*10),3] <- ws[i,1]
 data[(10*(i-1)+1):(i*10),4] <- ws[i,2]
 data[(10*(i-1)+1):(i*10),1] <- data[(10*(i-1)+1):(i*10),1] - mean(data[(10*(i-1)+1):(i*10),1])
 data[(10*(i-1)+1):(i*10),2] <- data[(10*(i-1)+1):(i*10),2] - mean(data[(10*(i-1)+1):(i*10),2])
}
r2MLMcomp(data,within_covs_modA=c(1,2),between_covs_modA=c(3),random_covs_modA=c(1,2),
     gamma_w_modA=c(2.5,-1),gamma_b_modA=c(1,4),Tau_modA=matrix(c(8,1,.75,1,1,.25,.75,.25,.5),3,3),sigma2_modA=10,
     within_covs_modB=c(1,2),between_covs_modB=c(3,4),random_covs_modB=c(1,2),
     gamma_w_modB=c(2.5,-1),gamma_b_modB=c(1,4,1),Tau_modB=matrix(c(6.5,.5,.25,.5,1,.25,.75,.25,.5),3,3),sigma2_modB=10)
```