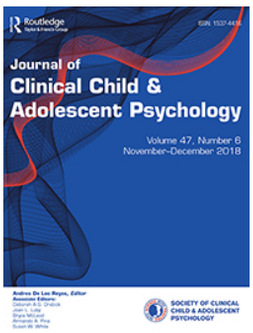


Journal of Clinical Child & Adolescent Psychology

ISSN: 1537-4416 (Print) 1537-4424 (Online) Journal homepage: https://www.tandfonline.com/loi/hcap20

Effect Size Measures for Multilevel Models in Clinical Child and Adolescent Research: New R-Squared Methods and Recommendations

Jason D. Rights & David A. Cole

To cite this article: Jason D. Rights & David A. Cole (2018) Effect Size Measures for Multilevel Models in Clinical Child and Adolescent Research: New R-Squared Methods and Recommendations, Journal of Clinical Child & Adolescent Psychology, 47:6, 863-873, DOI: 10.1080/15374416.2018.1528550

To link to this article: https://doi.org/10.1080/15374416.2018.1528550

Published online: 15 Nov 2018.

Journal of Clinical Child & Adolescent Psychology, 47(6), 863–873, 2018

ISSN: 1537-4416 print/1537-4424 online
DOI: https://doi.org/10.1080/15374416.2018.1528550

EVIDENCE BASE UPDATE

Effect Size Measures for Multilevel Models in Clinical Child and Adolescent Research: New R-Squared Methods and Recommendations

Jason D. Rights and David A. Cole
Department of Psychology and Human Development, Vanderbilt University

Clinical psychologists studying child and adolescent populations commonly analyze hierarchically structured data via multilevel modeling (MLM). In clinical child and adolescent psychology, and in psychology more broadly, increasing emphasis is being placed on the reporting of effect size, such as R-squared (R^2) measures of explained variance. In MLM, however, the literature on R^2 had, until recently, suffered from several shortcomings: (a) the relations among existing measures were unknown, (b) methods for quantifying some types of explained variance were unavailable, (c) which (if any) measures should be used for model comparison was unclear, (d) most measures did not generalize to models with more than two levels, and (e) software to compute measures was unavailable. The purpose of this article is to summarize recent methodological developments that resolved these issues and encourage the use of MLM R^2 in practice. We provide a nontechnical discussion of how the issues have been resolved and demonstrate how the new measures and methods can be implemented, highlighting their utility with an empirical example. We first consider a two-level MLM for a single hypothesized model in which we examine emotional response to social situations as a predictor of maladaptive self-cognitions, demonstrating the various ways we can quantify explained variance. We then discuss and demonstrate the use of R^2 for model comparison, and discuss the extension to models with more than two levels. Last, we discuss new free software that researchers can use to compute measures and produce associated graphics.

INTRODUCTION

Clinical psychologists studying child and adolescent populations commonly analyze hierarchically structured data via multilevel modeling (MLM).[1] Classic examples of such data include children nested with families, children nested within therapy groups, families nested within clinicians, students nested within classrooms, or repeated observations nested within persons (e.g., Ciesla, Reilly, Dickson, Emanuel, & Updegraff, 2012; Hagan et al., 2012; Salmivalli, Voeten, & Poskiparta, 2011). Sometimes, multiple levels of nesting are required (e.g., Hawley & Weisz, 2005). The MLM accounts for this nesting and enables researchers to model all hierarchical levels simultaneously.

In clinical child and adolescent psychology, and in psychology more broadly, there is increasing emphasis on reporting effect size, such as R-squared (R^2) measures of explained variance (American Psychological Association, 2008, 2009; Appelbaum et al., 2018; LaHuis, Hartman, Hakoyama, & Clark, 2014; Roberts, Monaco, Stovall, & Foster, 2011). Until recently, however, the MLM R^2 literature has suffered from several shortcomings. Rights and Sterba (2018a) noted five major issues:

1. The relations among the available measures have been unclear, causing concern about reconciling and interpreting differences among existing measures.

Correspondence should be addressed to Jason D. Rights, Quantitative Methods Program, Department of Psychology and Human Development, Vanderbilt University, Peabody #552, 230 Appleton Place, Nashville, TN 37203. E-mail: jason.d.rights@vanderbilt.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hcap.

[1] In the past 10 years, approximately 17% of articles published in *Journal of Clinical Child & Adolescent Psychology* have included analyses or discussion of MLM, also commonly called hierarchical linear modeling or mixed-effects modeling (estimated using Google Scholar and determining the percentage of articles using the exact phrase "multilevel modeling," its synonyms, and such variants as "multilevel models").

2. Existing measures have quantified only a few types of explained variance in MLM, limiting the kinds of substantive research questions that can be addressed.
3. Model comparisons are common in MLM; however, which R^2 measures should be used in the context of model comparison has been unclear.
4. Many measures do not generalize beyond two-level models.
5. Software to compute measures has not been readily available.

Consequently, MLM applications in child and adolescent clinical psychology often do not include estimates of R^2, or when they do, the reported R^2 might not correspond to the research question at hand. R^2s are especially rare in model comparison research and when models have more than two levels. This state of affairs contrasts with current practices in standard, single-level regression wherein researchers almost always report R^2 and conflicts with current reporting standards in the psychological sciences that emphasize effect size (American Psychological Association, 2008, 2009; Appelbaum et al., 2018).

The purpose of this article is to summarize recent developments spanning multiple methodological articles that have resolved the aforementioned issues (Rights & Sterba, 2018a, 2018b, 2018c). Our broader goal is to encourage the more appropriate use of MLM R^2 in practice. We provide a brief, nontechnical discussion of how the issues have been resolved and demonstrate how the new measures and methods can be implemented, highlighting their utility for clinical child and adolescent psychology applications with an empirical example. We begin by considering a two-level MLM for a single hypothesized model, then discuss and demonstrate the use of R^2 for model comparison, then discuss the extension to models with more than two levels, and finally make note of newly developed, free software that researchers can use for each of these contexts.

BACKGROUND: MLM

To begin, we introduce and describe an MLM specification that we use as an empirical example to facilitate discussion and demonstration of R^2 computation and interpretation. The model here is based on the cross-sectional analysis in Cole, Zelkowitz, Nick, Lubarsky, and Rights (in press), in which multiple observations are nested within adolescents; note, however, that the forthcoming R^2 measures and methods can similarly be applied to other nested data contexts, including longitudinal analyses (e.g., ecological momentary assessments) or clustering of persons into groups (e.g., children nested within therapy groups or students nested within classrooms). Of substantive interest in this empirical example is the degree to which adolescents' emotional responses to hypothetical social situations explain variation in maladaptive self-cognitions in response to said situations. Although numerous theories and studies have linked emotional responses to negative self-cognitions (e.g., Kovacs & Beck, 1978; Rosenberg, 1998), it is typically conceptualized as a between-subject phenomenon. In the current example, however, we are interested in quantifying separately the impact of the *within-person* and the *between-person* components of emotional responses. The within-person effect represents the degree to which *deviations in emotional response from one's own baseline* predicts maladaptive self-cognitions. A between-person effect represents the degree to which one's *baseline emotional response* predicts maladaptive self-cognitions.

Data were obtained via an adaptation of Cole et al.'s (2014) Behind Your Back procedure in which adolescents respond to 21 brief audio recordings. Participants were instructed to imagine the two people conversing in the recording are talking about them, with the content ranging from mild (e.g., "Do you want to let her work with us?" "Well, I guess I don't mind.") to mean (e.g., "He's so clueless!"). Emotional responses to each scenario were assessed by asking participants how sad (*SAD*) and how mad (*MAD*) they would feel (each on a 5-point Likert scale) in response to overhearing such a conversation. Cognitive responses (*COG*) were assessed by asking participants how much this would make them think negative thoughts about themselves (two items on 5-point Likert scales, averaged together).

In this design, the various scenarios (Level 1) are nested within adolescents (Level 2, or cluster level). To assess potential within-person (or within-cluster) effects of *SAD* and *MAD* on *COG*, we included in the model person-mean-centered *SAD* and *MAD*, which involves subtracting from each raw variable the person's mean of that variable across all 21 scenarios. To assess potential between-person (or between-cluster) effects, we included person-mean *SAD* and *MAD* as predictors of *COG*. The Level 1 (observation-level) model is given as

$$COG_{ij} = \beta_{0j} + \beta_{1j}(SAD_{ij} - SAD_{\cdot j}) + \beta_{2j}(MAD_{ij} - MAD_{\cdot j}) + e_{ij} \quad (1)$$

Here, i denotes scenario (one of the 21 audio recordings) and j denotes person (with $\cdot j$ subscripts indicating person-specific means). The Level 1 residual, e_{ij}, is normally distributed. In the Level 2 model, we define the person-specific regression intercept (β_{0j}) and slopes (β_{1j} and β_{2j}). Here we include the Level 2 predictors as well as the fixed and random effects, given by the γs and us, respectively:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}\mathrm{SAD}_{\cdot j} + \gamma_{02}\mathrm{MAD}_{\cdot j} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \\ \beta_{2j} &= \gamma_{20} + u_{2j} \end{aligned} \quad (2)$$

The person-specific intercepts are modeled by a fixed, across-person average intercept, γ_{00}, the between-person fixed effects of *SAD* and *MAD*, γ_{01} and γ_{02}, respectively, and a random component, that is, a person-specific residual, u_{0j}. The person-specific slopes are each modeled by a within-person fixed effect, γ_{10} or γ_{20}, as well a random person-specific residual, u_{1j} or u_{2j}. Together, the three Level 2 residuals are multivariate normally distributed. Note that in this analysis we are not interested in any cross-level interactions, which involve predicting the cluster-specific slopes with Level 2 variables. Such cross-level interaction terms (i.e., the product of a Level 1 and Level 2 variable), however, could be included in other models and could contribute to explained variance in the forthcoming R^2 measures.

From the model given in Equations 1 and 2, we can see that there are five entities that will influence the extent to which negative self-evaluations differ across observations: (a) the strength of the within-person fixed effects of emotional response, (b) the strength of between-person fixed effects of emotional response, (c) the degree of across-person variability in the effect of emotional response, (d) the degree of across-person variability in random intercepts, and (e) the degree of Level 1 residual variation. These five separate sources of variation are each involved in the computation of R^2 in the following section.

OVERVIEW OF FRAMEWORK OF MLM R^2s

An R^2 can be generically defined as the ratio of explained variance to the overall outcome variance, that is,

$$R^2 = \frac{\text{explained variance}}{\text{outcome variance}}. \quad (3)$$

This indicates the proportion of outcome variance (e.g., variation in *COG*) that can be explained by a given model, serving as a useful effect size with an intuitive metric (proportion) and logical bounds (0 and 1).

In comparison to single-level regression in which R^2 is routinely reported, multilevel R^2s are complicated by the fact that (a) there are multiple choices for what can be considered the outcome variance of interest (in the denominator) and (b) there are multiple sources that can be thought to contribute to explained variance (in the numerator). For instance, in quantifying variance in *COG* explained, a researcher might be interested in within-person differences (i.e., why adolescents would respond differently to different situations), between-person differences (i.e., why adolescents respond differently from one another), or both. Furthermore, of the sources of variation just listed, only certain ones might be of substantive interest.

Rather than providing a single R^2 to assess explained variance, Rights and Sterba (2018a) provided a suite of such measures that, together, provide complimentary information. This framework includes three distinct types of measures, differentiated by the outcome variance in the denominator—*within-cluster, between-cluster*, and *total* measures. The measures of each type are further distinguished by which of the available sources of variance are considered sources of explanation. This is accomplished by an analytic partitioning of the MLM-implied outcome variance (Rights & Sterba, 2018a). Specifically, the total outcome variance for an MLM is decomposed into five parts, that is, variance attributable to one of five sources: (a) Level 1 predictors via fixed slopes, (c) Level 2 predictors via fixed slopes, (c) Level 1 predictors via random slope variation, (d) cluster-specific outcome means via random intercept variation, and (e) Level 1 residuals. Note that these correspond (in order) to the five sources just listed in the context of the illustrative example. The first four of these have been deemed sources of *explained* variance in preexisting MLM R^2 measures. As a shorthand, we refer to these as f_1, f_2, v, and m, respectively.

Total measures have the total outcome variance in the denominator, which consists of variance attributable to all five sources just listed. *Within-cluster measures* include only variance attributable to f_1, v, and Level 1 residuals in the denominator, as together these comprise all of the within-cluster variance. Similarly, *between-cluster measures* include variance attributable to f_2 and m, yielding the between-cluster variance. The available measures are listed and defined in Table 1, with the subscript denoting the type of measure (total = t, within-cluster = w, between-cluster = b), and the superscript denoting the source(s) of explained variance.[2] The measures defined in Table 1 assume that all Level 1 predictors in the model are cluster-mean-centered, as this facilitates partitioning both explained variance and outcome variance into within-cluster versus between-cluster components. Nonetheless, five of the total measures —$R_t^{2(f)}$, $R_t^{2(v)}$, $R_t^{2(m)}$, $R_t^{2(fv)}$, and $R_t^{2(fvm)}$—can be computed even when Level 1 predictors are not cluster-mean-centered (for details, see Rights & Sterba, 2018a). Also note that, when using cluster-mean-centering, cross-level interaction terms explain variance exclusively within-cluster and thus contribute to f_1; when not cluster-mean-centering, cross-level interactions contribute to f (the sum of f_1 and f_2).

Although the sheer number of possible definitions may seem initially overwhelming, their use can be streamlined by a simple graphical representation (see Figure 1). This shows the decomposition of total, within-cluster, and between-cluster outcome variance into proportions, based on the aforementioned five sources (three of which are

[2] Note that some measures include only one source of explained variance, whereas others combine multiple sources. For simplicity, we focus on the single-source measures, as the combined-source measures are simple combinations of these, as noted in Table 1.

TABLE 1
Multilevel Model Only R^2 Measures in Integrative Framework for a Single Model in Isolation

Measure	*Definition (Interpretation)*
Total MLM R^2 measures	
$R_t^{2(f_1)}$	Proportion of total outcome variance explained by *Level 1 predictors via fixed slopes*
$R_t^{2(f_2)}$	Proportion of total outcome variance explained by *Level 2 predictors via fixed slopes*
$R_t^{2(v)}$	Proportion of total outcome variance explained by *Level 1 predictors via random slope variation/covariation*
$R_t^{2(m)}$	Proportion of total outcome variance explained by *cluster-specific outcome means via random intercept variation*
$R_t^{2(f)} = R_t^{2(f_1)} + R_t^{2(f_2)}$	Proportion of total outcome variance explained by *all predictors via fixed slopes*
$R_t^{2(fv)} = R_t^{2(f_1)} + R_t^{2(f_2)} + R_t^{2(v)}$	Proportion of total outcome variance explained by *predictors via fixed slopes and random slope variation/covariation*
$R_t^{2(fvm)} = R_t^{2(f_1)} + R_t^{2(f_2)} + R_t^{2(v)} + R_t^{2(m)}$	Proportion of total outcome variance explained by *predictors via fixed slopes and random slope variation/covariation* and by *cluster-specific outcome means via random intercept variation*
Within-cluster MLM R^2 measures	
$R_w^{2(f_1)}$	Proportion of within-cluster outcome variance explained by *Level 1 predictors via fixed slopes*
$R_w^{2(v)}$	Proportion of within-cluster outcome variance explained by *Level 1 predictors via random slope variation/covariation*
$R_w^{2(f_1v)} = R_w^{2(f_1)} + R_w^{2(v)}$	Proportion of within-cluster outcome variance explained by *Level 1 predictors via fixed slopes and random slope variation/covariation*
Between-cluster MLM R^2 measures	
$R_b^{2(f_2)}$	Proportion of between-cluster outcome variance explained by *Level 2 predictors via fixed slopes*
$R_b^{2(m)}$	Proportion of between-cluster outcome variance explained by *cluster-specific outcome means via random intercept variation*

Note: Measures presented here are assuming that all Level 1 predictors are cluster-mean-centered in the fitted multilevel modeling (MLM); however, the latter five total measures can also be computed for non-cluster-mean-centered models (for details, see Rights & Sterba, 2018a).

purely within-cluster; and two of which are purely between-cluster). These proportions form each of the R^2 measures. For instance, in this hypothetical example in Figure 1, it can be readily seen that f_1 explains about 25% of the total variance ($R_t^{2(f_1)} = .25$) and 40% of the within-cluster variance ($R_w^{2(f_1)} = .40$).[3] Researchers may have valid substantive reasons to prefer certain measures over others; for instance, an application may primarily seek to explain within-cluster variability rather than between-cluster. Nevertheless, they should still be interpreted in the context of the full decomposition illustrated by Figure 1 (see Rights & Sterba, 2018a).

The measures in Table 1 and Figure 1 comprise a unifying framework that subsumes preexisting measures. In Table 2, we list the authors that have developed MLM R^2 measures and use Xs to indicate which measures are equivalent (in the population[4]) to those shown in Table 1 (for details and derivations, see Rights & Sterba, 2018a). This integrative framework clarifies the relations among preexisting measures. Furthermore, this highlights the fact that several measures have been developed multiple times in the methodological literature, although their equivalencies have not been noted. In addition, Table 2 shows that these preexisting measures leave gaps in what researchers are able to quantify. For instance, none of these measures explicitly quantifies variance attributable to f_1 versus f_2. This distinction is important in the illustrative example.

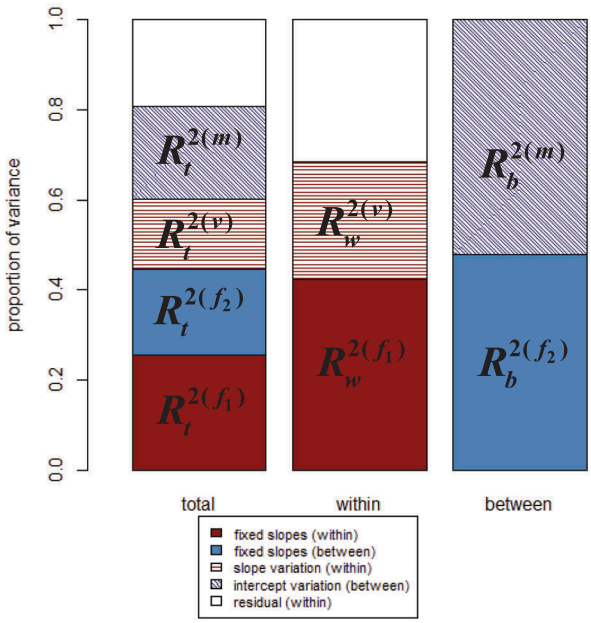
EMPIRICAL EXAMPLE: A SINGLE MODEL IN ISOLATION

We now examine the empirical example (using the model in Equations 1 and 2) to demonstrate the utility of the suite of R^2 measures. Conventional MLM results are presented in Table 3. The statistically significant estimates suggest that each of the four sources f_1, f_2, v, and m contribute to variation in *COG*. The *practical significance* of these, however, is difficult to quantify by merely examining the point estimates. For example, one may wonder: Do the Level 1 fixed slopes of .22 and .13 explain a sizable portion of the variance in *COG*? To address such questions, we use the R^2 measures listed in Table 1.

The R^2 estimates for this model appear in Figure 2 alongside an associated barchart displaying the decomposition of scaled outcome variance. First, we focus on the fixed effects. Results suggest that both within-person

[3] Any of the combined source measures can be visualized by looking at the cumulative size of the stacked bars; for instance, in the Figure 1 hypothetical example, f_1, f_2, v, and m together explain about 80% of the total outcome variance ($R_t^{2(fvm)} = .80$).

[4] Certain measures are additionally mathematically equivalent in the sample, as explained in Rights and Sterba (2018a).



Total R^2 measures

$$R_t^{2(f_1)}$$
$$R_t^{2(f_2)}$$
$$R_t^{2(v)}$$
$$R_t^{2(m)}$$
$$R_t^{2(f)} = R_t^{2(f_1)} + R_t^{2(f_2)}$$
$$R_t^{2(fv)} = R_t^{2(f_1)} + R_t^{2(f_2)} + R_t^{2(v)}$$
$$R_t^{2(fvm)} = R_t^{2(f_1)} + R_t^{2(f_2)} + R_t^{2(v)} + R_t^{2(m)}$$

Within-cluster R^2 measures

$$R_w^{2(f_1)}$$
$$R_w^{2(v)}$$
$$R_w^{2(f_1v)} = R_w^{2(f_1)} + R_w^{2(v)}$$

Between-cluster R^2 measures

$$R_b^{2(f_2)}$$
$$R_b^{2(m)}$$

FIGURE 1 Visualizing an integrative framework of R^2 measures: Decomposition of scaled total, within-cluster, and between-cluster outcome variance. *Note*: Each stacked bar reflects variance attributable to a single source. For instance, the stacked bar labeled $R_t^{2(f_1)}$ reflects the proportion of variance attributable (or explained by) Level 1 predictors via fixed slopes.

TABLE 2
Relations Among Measures in Rights and Sterba (2018a) Multilevel Modeling R^2 Framework and Previous Author's Measures

	Total Measures							*Within-Cluster Measures*		*Between-Cluster Measures*		
Authors	$R_t^{2(f_1)}$	$R_t^{2(f_2)}$	$R_t^{2(v)}$	$R_t^{2(m)}$	$R_t^{2(f)}$	$R_t^{2(fv)}$	$R_t^{2(fvm)}$	$R_w^{2(f_1)}$	$R_w^{2(v)}$	$R_w^{2(f_1v)}$	$R_b^{2(f_2)}$	$R_b^{2(m)}$
Vonesh and Chinchilli (1997)					X		X			X		
Snijders and Bosker (2012)					X							
Xu (2003)							X					
Aguinis and Culpepper (2015)			X									
Johnson (2014; extension of Nakagawa & Schielzeth, 2013)					X		X					
Raudenbush and Bryk (2002)										X	X	

Note: An X indicates that the authors (listed by row) developed a measure equivalent in the population to the Rights and Sterba (2018a) measures (listed by column).

and between-person fixed effects of emotional response are of practical importance. The former explains 10% ($\hat{R}_t^{2(f_1)} = .10$) of the total variance and 22% ($\hat{R}_w^{2(f_1)} = .22$) of the within-person variance in *COG*—an adolescent's scenario-specific degree of negative emotional response relative to baseline emotional response may play a meaningful role[5] in determining the extent of maladaptive self-cognitions. The latter explains 18% ($\hat{R}_t^{2(f_2)} = .18$) of the total variance and 35% ($\hat{R}_b^{2(f_2)} = .35$) of the between-person variance—an adolescent's baseline level of emotional response also has a substantial relation to maladaptive self-cognitions. These results suggest that researchers studying potential links between emotional response and

[5] Researchers may wish to employ Cohen's (1992) rules-of-thumb to determine whether an effect size is small ($R^2 = .02$), medium ($R^2 = .13$), or large ($R^2 = .26$). We caution researchers, however, that such guidelines are to some degree subjective, arbitrary, and dependent on the research context.

TABLE 3
Empirical Example: Parameter Estimates for a Single Model in Isolation

Fixed effects	*Estimate*	SE	t
Intercept	0.39	0.13	3.03*
Person-mean-centered SAD	0.22	0.01	17.40*
Person-mean-centered MAD	0.13	0.01	12.32*
Person-mean SAD	0.57	0.04	15.01*
Person-mean MAD	0.08	0.05	1.90
Variance components	*Estimate*	SE	z
Variance of intercept	0.62	0.04	15.96*
Variance of person-mean-centered *SAD*	0.04	0.01	7.90*
Variance of person-mean-centered *MAD*	0.02	< 0.01	6.45*
Covariance of intercept with person-mean-centered *SAD*	0.06	0.01	5.61*
Covariance of intercept with person-mean-centered *MAD*	0.07	0.01	6.88*
Covariance of person-mean-centered *SAD* with person-mean-centered *MAD*	< 0.01	< 0.01	1.22
Variance of Level 1 residuals	0.54	0.01	71.89*

Note: Results were obtained via SPSS using the MIXED command and maximum likelihood estimation.

*$p < .05$ (significance of random components was assessed using the alpha correction approach of Fitzmaurice, Laird, & Ware, 2011).

negative self-cognition should be careful to consider *both* within-person and between-person effects; if we had assessed only the latter (as is commonly done) we would essentially neglect the estimated 10% of total variance and 22% of within-person variance that can be explained when also considering within-person effects.

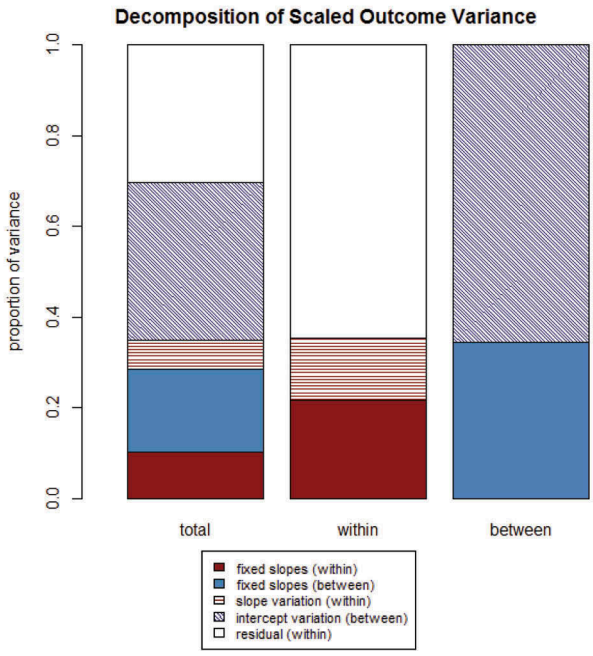
We next consider the importance of random effect variation. Results suggest that individual differences in the effect of emotional response are important—the random slope variation accounts for 6% of the total variance ($\hat{R}_t^{2(v)} = .06$) and 14% of the within-person variance ($\hat{R}_w^{2(v)} = .14$). Although researchers may be more accustomed to ascribing substantive meaning to the fixed portion of slopes, in this application, the random slope variance is also important. The association of emotional response to cognitive response is stronger for some adolescents than others; Cole et al. (in press) described this as individual differences in adolescents' *cognitive reactivity* to negative events. Given that this accounts for a sizable portion of the outcome variance, researchers studying the link between emotional and cognitive reactions would be advised to account for it in both study design and analysis. As for the influence of random intercept variation ($\hat{R}_t^{2(m)} = .35$ and $\hat{R}_b^{2(m)} = .65$), this is not a substantive focus in this application, but it is nonetheless useful to see the extent of individual differences in *COG* above that accounted for by the predictors in the model. The large values indicate that there is still a lot that differentiates adolescents, suggesting that future modeling could include additional adolescent-level (Level 2) predictors to account for these differences.

MLM R^2s FOR MODEL COMPARISON

We now turn our attention to the use of R^2 *differences*, or ΔR^2, in the context of model comparison. In single-level contexts, researchers routinely add terms to a given model and compute the difference between the R^2 from the full model and that of the reduced model. This is useful in assessing the practical importance of added terms, that is, how much variance they explain above and beyond the terms in the reduced model. As mentioned earlier, however, this is rarely done for MLM. Beyond the general R^2 issues just discussed, some additional issues have prevented ΔR^2 specifically. In particular, methodologists have shown via simulation that existing measures are unable to detect meaningful differences between models (e.g., Jaeger, Edwards, Das, & Sen, 2017; Orelien & Edwards, 2008). However, as discussed in Rights and Sterba (2018c), these comparisons were made using measures that implicitly combine multiple sources of either explained or unexplained variance in ways that made them unsuited to detect such differences in the first place. For example, authors have stated that $\Delta R_t^{2(f)}$ analogs have limited utility in that they are unable to detect the addition of a random component of a slope; however, by definition, $R_t^{2(f)}$ reflects variance explained by fixed effects only, so it is not designed to detect such an addition in the first place. $R_t^{2(v)}$, in contrast, is capable of detecting such an addition.

Rights and Sterba (2018a, 2018c) address this issue by outlining a step-by-step procedure to find a *match* between the model comparison being made and the relevant R^2 to consider, outlined in Table 4. The first step is to determine the term(s) that will be added to Model A (the reduced model) to form Model B.[6] For instance, to assess the practical importance of a Level 1 predictor via a fixed slope, one can add this term to Model A. Note that one can also add multiple terms; provided there is only one of each *type* of term added, one can assess the variance explained uniquely by each term. If multiple terms of the *same* type are added, one can assess only the joint impact of these. Step 2 is to determine whether interest is in assessing level-specific variance explained or total variance explained or both. Step 3 is to compute and focus interpretation on the measure corresponding to the choices in Steps 1 and 2. The final Step 4 is to

[6] In this table, we assume that the random intercept will be added only to a Model A that contains a fixed intercept and no predictors. This can be used as an initial comparison to assess the extent of overall between-cluster variability. However, this first step can equivalently be accomplished by computing the intraclass correlation coefficient (i.e., the ratio of between-cluster variance to the sum of between- and within-cluster variances) from a random-intercept-only model.



Total R^2 measures

$$\hat{R}_t^{2(f_1)} = .10$$
$$\hat{R}_t^{2(f_2)} = .18$$
$$\hat{R}_t^{2(v)} = .06$$
$$\hat{R}_t^{2(m)} = .35$$
$$\hat{R}_t^{2(f)} = \hat{R}_t^{2(f_1)} + \hat{R}_t^{2(f_2)} = .29$$
$$\hat{R}_t^{2(fv)} = \hat{R}_t^{2(f_1)} + \hat{R}_t^{2(f_2)} + \hat{R}_t^{2(v)} = .35$$
$$\hat{R}_t^{2(fvm)} = \hat{R}_t^{2(f_1)} + \hat{R}_t^{2(f_2)} + \hat{R}_t^{2(v)} + \hat{R}_t^{2(m)} = .70$$

Within-cluster R^2 measures

$$\hat{R}_w^{2(f_1)} = .22$$
$$\hat{R}_w^{2(v)} = .14$$
$$\hat{R}_w^{2(f_1v)} = \hat{R}_w^{2(f_1)} + \hat{R}_w^{2(v)} = .35$$

Between-cluster R^2 measures

$$\hat{R}_b^{2(f_2)} = .35$$
$$\hat{R}_b^{2(m)} = .65$$

FIGURE 2 Empirical example: R^2 estimates for a single model in isolation.

visualize and interpret the target ΔR^2 in the context of the set of all single-source measures, which can be done easily with a barchart graphic, as we demonstrate in the next section.[7] With this procedure, for instance, researchers can compare a model with random slopes to one without random slopes and, by following steps in Table 4, use $\Delta R_t^{2(v)}$ (as opposed to $\Delta R_t^{2(f)}$, which is irrelevant in this context) to determine the importance of random slopes in terms of variance explained.[8]

RUNNING EMPIRICAL EXAMPLE: COMPARING TWO MODELS

In the model in Equations 1 and 2, we included *SAD* and *MAD* in the model simultaneously to together represent emotional response. One may, however, be specifically interested in the unique contribution of sadness, as it is closely tied with the development of depressive schema (Cole et al., in press; Kovacs & Beck, 1978).

To illustrate, we consider four models. Model A is a simple fixed intercept null (i.e., without predictors) model. Model B adds to Model A a random component to the intercept, making a random-intercept-only model, which allows us to assess the overall proportion of variance that is between-person. Model C adds to Model B a random slope (i.e., both a fixed and random component) of person-mean-centered *MAD* and a fixed slope of person-mean *MAD*, whereas Model D adds the same terms for *SAD* (yielding Equations 1 and 2). Adding *MAD* and *SAD* separately in this way allows us to determine the importance of *SAD* over and above *MAD*.

In Figure 3, we present results for each of the three model comparisons, showing the decomposition of scaled total variance for each model side by side in the top row; we did the same for the within-cluster variance in row 2 and between-cluster variance in row 3. With reference to the procedure from Table 4, we first consider the *B* to *A* comparison. Here we are adding a random intercept, and thus the relevant measure is $\Delta R_t^{2(m)}$. Here we estimate this to be .54 ($\Delta\hat{R}_t^{2(m)} = .54$ listed below the bar for Model B), meaning an estimated 54% of the variability in *COG* is attributable to between-person differences (highlighting the need for MLM). For the *C* to *B* comparison, we can see the contribution of *MAD* above and beyond the random-intercept-only model; note that we add each of the

[7] As an optional Step 5, one can compute a combined-source ΔR^2 by simply adding the relevant ΔR^2 measures; we recommend, however, that researchers focus on the single-source measures for more complete information (see Rights & Sterba, 2018c, for details).

[8] We caution researchers, however, in using this information alone in determining whether random slopes are necessary to include. In general, a low ΔR^2 simply reflects a small effect size and does not necessarily imply that the added terms are superfluous (see Rights & Sterba, 2018a).

TABLE 4
Step-by-Step Procedure for Model Comparison Using ΔR^2 Measures

Step	*Choose Appropriate Column(s) at Each Step to Identify Measures to Interpret*							
Step 1: What term(s) does Model B add to Model A?	Random Intercept[a]		Fixed Component of a Level 1 Predictor[b]		Random Component of a Level 1 Predictor		Fixed Component of a Level 2 Predictor	
Step 2: Is interest in quantifying the impact of added term(s) relative to total variance, level-specific variance, or both?	Total measure	Level 2 measure	Total measure	Level 1 measure	Total measure	Level 1 measure	Total measure	Level 2 measure
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[c]	$\Delta R_t^{2(f_1)}$	$\Delta R_w^{2(f_1)}$	$\Delta R_t^{2(v)}$	$\Delta R_w^{2(v)}$	$\Delta R_t^{2(f_2)}$	$\Delta R_b^{2(f_2)}$
Step 4: Visualize and interpret changes in target single-source measure(s) in the context of the set of all single-source measures.	See Figure 3 example							

Note: Procedure described here assumes that all Level 1 predictors are cluster-mean-centered in both fitted MLMs; see Rights and Sterba (2018a) for detail on computation for non-cluster-mean-centered models.
[a]Here we assume that the random intercept will only be added to a Model A that contains only a fixed intercept and no predictors. This can be used as an initial comparison to assess the extent of overall between-cluster variability. However, this first step can equivalently be accomplished by computing the intraclass correlation coefficient (i.e., the ratio of between-cluster variance to the sum of between- and within-cluster variances) from a random-intercept-only model.
[b]Cross-level interaction terms exclusively explain within-cluster variance when Level 1 predictors are cluster-mean-centered (Rights & Sterba, 2018a), and thus adding one can be considered adding a "fixed component of a Level 1 predictor." Cross-level interactions can also yield a decrease in $R_t^{2(v)}$ and $R_w^{2(v)}$, as they can account for across-cluster differences in slopes (Raudenbush & Bryk, 2002).
[c]$\Delta R_b^{2(m)}$ will necessarily be 1 when comparing a fixed intercept null model to a random intercept null model.

terms listed in Step 1 in Table 4 (aside from the already-included random intercept), and in Step 2 we determine that total variance is of primary interest (although we also present the level-specific measures and visualization for completeness); in Step 3, we thus quantify the contribution relative to the total variance by examining $\Delta R_t^{2(f_1)}$, $\Delta R_t^{2(v)}$, and $\Delta R_t^{2(f_2)}$, visualizing this change (Step 4) in Figure 3. We see that *MAD* alone explains a modest amount of the total variance by each of its sources: $\Delta\hat{R}_t^{2(f_1)} = .08$, $\Delta\hat{R}_t^{2(v)} = .06$, and $\Delta\hat{R}_t^{2(f_2)} = .06$. Note that the negative value for $\Delta\hat{R}_t^{2(m)}$ indicates that some of the intercept variance in Model B is now accounted for (or explained by) person-mean *MAD* in Model C; thus $\Delta\hat{R}_t^{2(m)} = -.06$ is equal in magnitude but opposite in sign to $\Delta\hat{R}_t^{2(f_2)} = .06$. In the final model comparison, D to C, we can see the contribution of *SAD* over and above *MAD*. Following the same steps as the previous comparison, we see that little variance is explained by the within-person component of *SAD* ($\Delta\hat{R}_t^{2(f_1)} = .02$ and $\Delta\hat{R}_t^{2(v)} = .02$), but a sizable portion is explained by the between-person component ($\Delta\hat{R}_t^{2(f_2)} = .12$).

The ΔR^2 results suggest there to be differential predictive utility of *SAD* at the within-person level and the between-person level. It could be the case that, for a given adolescent responding to different social situations, general negative emotional response—as opposed to either *MAD* or *SAD* uniquely—is predictive of maladaptive cognitions. On the contrary, at the between-person level, *SAD* uniquely differentiates adolescents, meaning that one's baseline level of sadness in response to social situations is an important predictive factor of maladaptive cognition. Such information could be particularly useful, for instance, when considering which factors are most important to target via intervention (of course, this would require further investigation in future research). The general point here is that the use of these ΔR^2 helps quantify the unique importance of predictors included in the model in a way that examining R^2 for a single model in isolation, or that using previously developed ΔR^2 procedures, does not allow.

EXTENSION: MODELS WITH THREE OR MORE LEVELS

Computation of R^2 is particularly rare in multilevel contexts with more than two levels, although examples of such data are not too uncommon in child and adolescent clinical psychology (e.g., Hawley & Weisz, 2005). One reason for this is that many existing measures simply do not extend past two levels. A further reason may be that, similar to what was just discussed in the context of two-level models, measures that exist cover only a limited subset of the possible ways to quantify explained

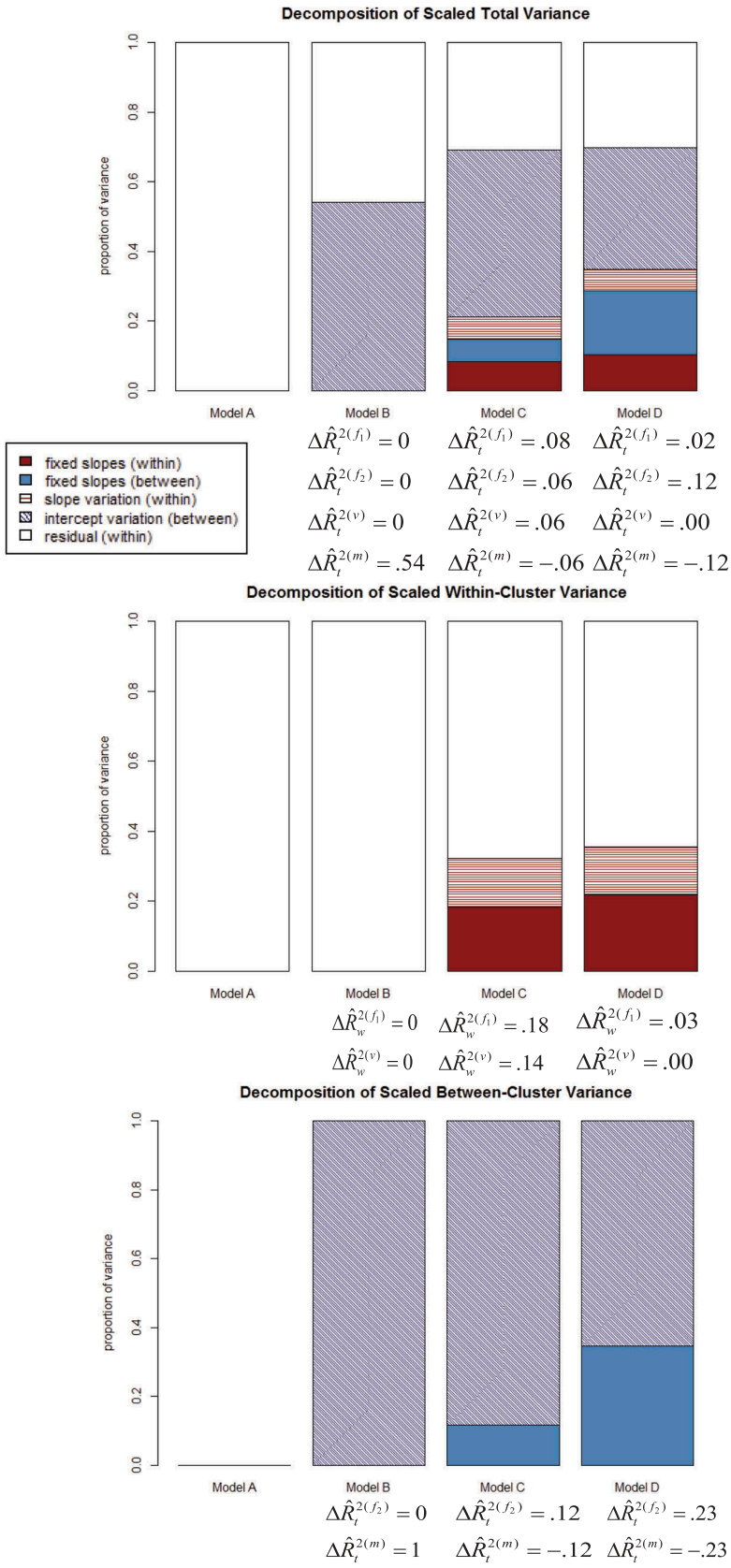


FIGURE 3 Empirical example: ΔR^2 estimates for three separate model comparisons.

variance. This is particularly problematic in higher level contexts because measures that implicitly combine sources of explained variance can do so across many different levels of a hierarchy simultaneously, thus making it unclear which of the potential sources across the levels are most important. For instance, if one were to report, say, a four-level $R_t^{2(f)}$ analog measure, this would reflect the cumulative variance explained by fixed effects at each of the four levels. If substantive interest were primarily in sources at a subset of these levels, a high $R_t^{2(f)}$ could be misleading; it could be the case that only the fixed effects at less substantively important levels are accounting for the majority of the explained variance. To address this, Rights and Sterba (2018b) extended their general framework developed for two-level models to accommodate any number of levels. With these measures, one can separately assess, for instance, the variance explained by Level 1 versus Level 2 versus Level 3 versus Level 4 predictors via fixed slopes.

With the current empirical example, a three-level analysis is not relevant. Although adolescents were nested within classrooms (a potential third level), classroom membership was unrelated to the outcome of interest, and there was a negligible degree of across-classroom variance in *COG*. As an illustrative example, however, suppose the third-level clustering was more relevant to *COG*, for instance, if adolescents were nested within clinicians. If certain clinicians were more effective in providing strategies to inhibit maladaptive cognition in response to social situations, then there could be substantial across-clinician variability in *COG*. We could then use the framework of Rights and Sterba (2018b) to quantify the importance of observation-level, adolescent-level, and clinician-level predictors, as well as random effect variation across adolescents and clinicians, and furthermore could quantify these with regards to total or level-specific variance.

SOFTWARE

To aid researchers in computing all the R^2 measures discussed here, three separate *R* functions have been developed: *r2MLM*, which computes the measures for a single model in isolation and produces the associated barchart (Rights & Sterba, 2018a; see supplemental materials at http://supp.apa.org/psycarticles/supplemental/met0000184/Supplemental_Materials); *r2MLMcomp*, which computes the ΔR^2 between two models and provides the associated barchart comparison (Rights & Sterba, 2018c); and *r2MLM3*, which computes all measures relevant for a three-level model and provides a barchart graphic[9] (Rights & Sterba, 2018b). Each of the functions, as well as more thorough descriptions, can be found at the Software page of the first author's webpage (https://my.vanderbilt.edu/jasonrights/software). At this webpage, we additionally provide example R code to simulate data and run each of the functions; researchers can modify this code to read in their own data and compute R^2 measures.

[9] Although the code currently does not accommodate more than three levels, researchers dealing with four or more levels (though rare) can compute measures using the formulae provided by Rights and Sterba (2018b).

CONCLUSION

By providing an overview of several methodological papers on MLM R^2 and by illustrating the use of such measures for clinical child and adolescent psychology, we hope that this article encourages researchers to assess explained variance in their MLM applications. Such a shift would be in line with calls to cease overreliance on statistical significance and to consider practical significance and effect size.

ACKNOWLEDGMENTS

We thank Sonya Sterba for helpful comments.

REFERENCES

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

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:10.1037/0003-066X.63.9.839

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

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*, 3–25. doi:10.1037/amp0000191

Ciesla, J. A., Reilly, L. C., Dickson, K. S., Emanuel, A. S., & Updegraff, J. A. (2012). Dispositional mindfulness moderates the effects of stress among adolescents: Rumination as a mediator. *Journal of Clinical Child & Adolescent Psychology*, *41*, 760–770. doi:10.1080/15374416.2012.698724

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cole, D. A., Martin, N. C., Sterba, S. K., Sinclair-McBride, K., Roeder, K. M., Zelkowitz, R., & Bilsky, S. A. (2014). Peer victimization (and harsh parenting) as developmental correlates of cognitive reactivity, a diathesis for depression. *Journal of Abnormal Psychology*, *123*, 336–349. doi:10.1037/a0036489

Cole, D. A., Zelkowitz, R. L., Nick, E. A., Lubarsky, S. R., & Rights, J. D. (in press). Simultaneously examining negative appraisals, emotion reactivity, and cognitive reactivity in relation to depressive symptoms in children. *Development and Psychopathology*.

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

Hagan, M. J., Tein, J. Y., Sandler, I. N., Wolchik, S. A., Ayers, T. S., & Luecken, L. J. (2012). Strengthening effective parenting practices over the long term: Effects of a preventive intervention for parentally bereaved families. *Journal of Clinical Child & Adolescent Psychology*, *41*, 177–188. doi:10.1080/15374416.2012.651996

Hawley, K. M., & Weisz, J. R. (2005). Youth versus parent working alliance in usual clinical care: Distinctive associations with retention, satisfaction, and treatment outcome. *Journal of Clinical Child and Adolescent Psychology, 34*, 117–128. doi:10.1207/s15374424jccp3401_11

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

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's $R2_{GLMM}$ to random slopes models. *Methods in Ecology and Evolution*, *5*, 944–946. doi:10.1111/2041-210X.12225

Kovacs, M., & Beck, A. T. (1978). Maladaptive cognitive structures in depression. *American Journal of Psychiatry*, *135*, 525–533. doi:10.1176/ajp.135.5.525

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods*, *17*, 433–451. doi:10.1177/1094428114541701

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

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

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

Rights, J. D., & Sterba, S. K. (2018a). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*. doi:10.1037/met0000184

Rights, J. D., & Sterba, S. K. (2018b). Generalizing a framework of multilevel model R-squared measures to accommodate three or more levels. Manuscript in preparation.

Rights, J. D., & Sterba, S. K. (2018c). New recommendations on the use of R-squared differences in multilevel model comparisons. Manuscript under review.

Roberts, J. K., Monaco, J. P., Stovall, H., & Foster, V. (2011). Explained variance in multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 219–230). New York, NY: Routledge.

Rosenberg, E. L. (1998). Levels of analysis and the organization of affect. *Review of General Psychology*, *2*, 247–270. doi:10.1037/1089-2680.2.3.247

Salmivalli, C., Voeten, M., & Poskiparta, E. (2011). Bystanders matter: Associations between reinforcing, defending, and the frequency of bullying behavior in classrooms. *Journal of Clinical Child & Adolescent Psychology*, *40*, 668–676. doi:10.1080/15374416.2011.597090

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

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

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