Analyzing Multiple Outcomes in Clinical Research Using Multivariate Multilevel Models

Scott A. Baldwin,
Brigham Young University

Zac E. Imel,
University of Utah

Scott R. Braithwaite, and
Brigham Young University

David C. Atkins
University of Washington

Abstract

Objective—Multilevel models have become a standard data analysis approach in intervention research. Although the vast majority of intervention studies involve multiple outcome measures, few studies use multivariate analysis methods. The authors discuss multivariate extensions to the multilevel model that can be used by psychotherapy researchers.

Method and Results—Using simulated longitudinal treatment data, the authors show how multivariate models extend common univariate growth models and how the multivariate model can be used to examine multivariate hypotheses involving fixed effects (e.g., does the size of the treatment effect differ across outcomes?) and random effects (e.g., is change in one outcome related to change in the other?). An online supplemental appendix provides annotated computer code and simulated example data for implementing a multivariate model.

Conclusions—Multivariate multilevel models are flexible, powerful models that can enhance clinical research.

Keywords

intervention data; multilevel; multivariate

Analyzing Multiple Outcomes in Clinical Research Using Multivariate Multilevel Models
Multilevel (mixed) models (Hox, 2010; Raudenbush & Bryk, 2002) have become a standard method for analyzing psychotherapy outcome data given the hierarchical structure of psychotherapy data, for example: (a) observations (level-1) clustered within persons (level-2) in longitudinal data (Singer & Willet, 2003), (b) patients (level-1) clustered within

Correspondence concerning this article should be addressed to: Scott A. Baldwin, 268 TLRB, Brigham Young University, Provo, UT 84602. scott_baldwin@byu.edu.
Scott A. Baldwin and Scott R. Braithwaite, Department of Psychology, Brigham Young University; Zac E. Imel, Department of Educational Psychology, University of Utah; David C. Atkins, Department of Psychiatry and Behavioral Sciences, University of Washington.

therapists or groups (level-2; Crits-Christoph & Mintz, 1991; Wampold & Serlin, 2000), and (c) effect sizes (level-1) clustered within studies (level-2) in meta-analysis (Hox, 2010). Moreover, it is not unusual to have additional grouping factors that can lead to three or more levels within psychotherapy data (e.g., repeated measures on individuals nested within couples; Atkins, 2005).

The hierarchical structure of psychotherapy data is important for both substantive and methodological reasons. Substantively, we are often interested in variability among higher-level factors, such as person-to-person variability or therapist-to-therapist variability (Baldwin & Imel, 2013; Crits-Christoph et al., 1991; Imel, Baldwin, Bonus, & Maccoon, 2008; Saxon & Barkham, 2012). Methodologically, the hierarchical structure leads to correlations among the observations within a cluster. For example, in group therapy, data from patients within the same group are likely to be correlated relative to data from patients in other groups. This correlation violates the assumption of independence of observations common to most statistical tests, which can lead to biased *p*-values, incorrect confidence intervals, and inflated effect sizes (Baldwin, Murray, & Shadish, 2005; Crits-Christoph & Mintz, 1991; Wampold & Serlin, 2000). Multilevel models accommodate the correlation among observations by modeling between-cluster variability via additional error terms called random effects (Raudenbush & Bryk, 2002; Singer & Willet, 2003) and are also more flexible regarding ignorable missing data and the correlation structure of the residuals than earlier methods (e.g., ANOVA; Singer & Willet, 2003). In addition, multilevel models can accommodate longitudinal data where participants are measured on different schedules (Hox, 2010), treat time as categorical, continuous, or some combination (Singer & Willet, 2003), be extended to situations where clustering affects some participants but not others (Baldwin, Bauer, Stice, & Rohde, 2011; Bauer, Sterba, & Hallfors, 2008), and accommodate non-normal outcomes (Rabe-Hesketh & Skrondal, 2008). Overall, multilevel models represent highly useful statistical tools for psychotherapy researchers.

However, one particular class of multilevel models has not been widely adopted within the psychotherapy research community—multivariate multilevel models, which extend multilevel models to two or more outcomes (Hox, 2010; MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997).[1] It is unusual to find psychotherapy research studies that only involve a single outcome variable. However, few researchers employ multivariate techniques when evaluating multiple outcomes, especially when multilevel models were used to analyze outcomes. As we will show, this omission results in a failure to test important theoretical questions that are best examined in the multivariate, multilevel context.

Historically, psychotherapy researchers regularly used multivariate data analysis methods, such as MANOVA, to control the experiment-wise Type I error rates. For example, a common strategy was to test for a treatment effect across outcomes using MANOVA and follow-up with series of ANOVA analyses on each outcome. Thus, even when MANOVA was used, the fundamental focus of the analysis was on univariate hypotheses. However, multivariate models can be used to address multivariate hypotheses. Examples of

---

[1]Occasionally, the term multivariate is used to refer to models with a single outcome variable but with multiple predictor variables. In this paper, we use the term multivariate to refer to models with two or more outcome variables.

multivariate hypotheses include testing whether outcomes have different average rates of change (Kaysen et al., 2011) or whether change in one outcome is related to change in another (Suvak, Walling, Iverson, Taft, & Resick, 2009).

The present paper introduces multivariate multilevel models for intervention research and illustrates how to fit and interpret the models. Example data throughout focuses on examining relationships between primary and secondary outcomes in a randomized trial. We examine whether predictor variables (e.g., treatment condition) have different relationships across outcomes and model relationships between outcomes (e.g., how is change in one outcome related to change in another). Although a brief overview of univariate multilevel models included as a bridge for to multivariate models, we assume that readers have a basic familiarity with univariate multilevel models. Readers desiring a more in-depth discussion of univariate models can consult a number of excellent textbooks (Gelman & Hill, 2007; Raudenbush & Bryk, 2002; Singer & Willet, 2003). We have included an online appendix that provides annotated code for fitting these models in Stata, SAS, SPSS, R, and Mplus.

Differential Fixed Effects Across Outcomes

Most intervention studies include multiple outcomes. For example, a study comparing cognitive-behavioral therapy and acceptance and commitment therapy for depression may include depression measures as well as quality of life measures, which could be classified as primary (depression) or secondary (quality of life) outcomes. Researchers may be interested in whether treatment effects differ by outcome type for both clinical and theoretical reasons. Clinically, patients presenting for treatment of depression may be most keenly interested in the impact of treatment on their depressive symptoms, though improvement in all areas of their life would likely be welcome. Further, a cognitive-behavioral therapy that explicitly targets depression symptoms might be expected to have an earlier and potentially larger impact on these symptoms (i.e., quality of life might improve as a result of decreases in depression). Theoretically, some have argued that larger treatment effects for primary outcomes are consistent with the hypothesis that factors specific to a treatment package, rather than factors common across treatment packages, are partially responsible for change (Hofmann & Lohr, 2010).

Most researchers do not directly test for differential treatment effects across outcomes, yet they often interpret their results as if they had tested for differential effects. Indeed, in our review work, we often find researchers invoking the "eyeball test" in such situations, noting simply that effects for primary outcomes were "larger" than those for secondary outcomes. Thus, results sections from clinical research usually involve using multilevel models, or a similar analysis technique, to examine intervention effects one outcome at a time. Even if the outcomes are grouped in primary and secondary categories, such grouping is usually for the purposes of the written report rather than incorporated directly into the data analysis. The results are then summarized with respect to which outcomes had significant treatment effects and which did not. For example, McDonagh et al. (2005) randomized participants meeting criteria for posttraumatic stress disorder (PTSD) to cognitive-behavioral therapy (CBT), present centered therapy (PCT), and wait-list (WL). In the analysis, the authors examined treatment effects across each outcome (using the group×time interaction in a

repeated measures ANOVA and post-hoc tests). The authors found a statistically significant intervention effect for PTSD symptoms but not for depression. The authors interpret the statistical significance for one outcome but not the other as important: "The fact that this treatment [PCT] had more of an impact on PTSD symptoms than on depressive symptoms suggests its mechanism is not simply an antidepressant effect such as has already been demonstrated for problem-solving therapy" (p. 522).

However, comparisons of intervention effects require a statistical test of the difference (cf. Nieuwenhuis, Forstmann, & Wagenmakers, 2011). In other words, we need an explicit test of whether the size of the intervention effect depends upon the outcome type—we need to test the intervention effect by outcome interaction. The idea is similar to moderator analyses in psychotherapy meta-analyses, in which one tests whether an effect size (i.e., treatment effect) varies as a function of study characteristics.

To get a sense of how often researchers analyze multiple outcomes and do not investigate whether treatment effects vary across outcomes, we reviewed randomized trials that used multilevel models published in the *Journal of Consulting and Clinical Psychology* from 2009 to 2012. We coded whether the study included multiple outcomes and whether the authors used multivariate methods to investigate differential treatment effects. We identified 60 randomized trials that used multilevel modeling to estimate treatment effects for multiple outcomes. Of these 60, one tested for differential treatment effects across outcomes (Jouriles et al., 2009), suggesting a profound mismatch between study design—with multiple outcomes—and study analysis.

Extending the Univariate Multilevel Model

Testing multivariate hypotheses about fixed effects can be accomplished by extending multilevel models to accommodate two or more outcomes. To illustrate the extension, we simulated data to mimic a clinical trial comparing cognitive-behavioral therapy (CBT) to a no-treatment control for the treatment of depression, three timepoints (baseline, midtreatment, posttreatment), 100 participants (50 per condition), and two outcomes— depression and quality of life. We coded time as 0, 1, and 2, with 0 representing the baseline timepoint. We also coded treatment condition (Tx) as 1 for CBT and 0 for control. In the population model, the treatment effect for depression was a .5 standard deviation difference at posttreatment (time 2) between CBT and control and there was no treatment effect for quality of life.

The univariate growth-curve models for each outcome can be written as follows (Singer & Willet, 2003):

$$y_{1ij} = \beta_{10} + \beta_{11} Time_{ij} + \beta_{12} Tx_j + \beta_{13} Time_{ij} Tx_j + u_{1j} + v_{1j} Time_{ij} + e_{1ij}, \quad (1)$$

$$y_{2ij} = \beta_{20} + \beta_{21} Time_{ij} + \beta_{22} Tx_j + \beta_{23} Time_{ij} Tx_j + u_{2j} + v_{2j} Time_{ij} + e_{2ij}. \quad (2)$$

Focusing on Equation (1), y_{1ij} is the depression outcome at time i for person j. β_{10} is the overall intercept, and like all intercepts, it represents the expected depression value (i.e.,

mean) of the outcome when all predictors are equal to 0, in this case when $Time_{ij} = 0$ (Baseline) and $Tx_j = 0$ (control). Changing the coding method for time or treatment condition (or by including other variables in the model) will alter the specific interpretation of the intercept (see Singer and Willet, 2003, for a discussion of alternative methods for coding time). β_{11} is the average rate of change in depression symptoms during treatment for the control condition, β_{12} is the mean difference between CBT and control at baseline, β_{13} is the difference in rate of change between CBT and control (i.e., the treatment effect), u_{1j} is a random effect representing person-specific differences at baseline (i.e., unique baseline values for each participant), v_{1j} is a random effect representing person-specific differences in change during treatment (i.e., unique rate of change for each participant), and e_{1ij} is residual error. The parameters in Equation (2) have identical interpretations except they pertain to quality of life.

Because the data in our example are longitudinal, the repeated observations within an individual are correlated. The random effects described above allow us to accommodate this correlation. Specifically, the models in Equations (1) and (2) assume that observations are independent conditional on the random effects (i.e., uncorrelated once the random effects are taken into account) and that the random effects are normally distributed (Singer & Willet, 2003):

$$\begin{bmatrix} u_{1j} \\ v_{1j} \end{bmatrix} \sim MVN \left(0, \begin{bmatrix} \sigma_{u1}^2 & \\ \sigma_{u1v1} & \sigma_{v1}^2 \end{bmatrix} \right), \quad (3)$$

$$\begin{bmatrix} u_{2j} \\ v_{2j} \end{bmatrix} \sim MVN \left(0, \begin{bmatrix} \sigma_{u2}^2 & \\ \sigma_{u2v2} & \sigma_{v2}^2 \end{bmatrix} \right). \quad (4)$$

Focusing on Equation (3) the random effects for depression, u_{1j} and v_{1j}, come from a normal distribution with a mean of 0, variances of σ_{u1}^2 and σ_{v1}^2, and a covariance of σ_{u1v1}^2. The parameters in Equation (4) have identical interpretations except that they pertain to quality of life. It may not be obvious at first glance, but Equations (3) and (4) encapsulate one of the advantages of multilevel models. The models shown in Equations (1) and (2) model the person-to-person variability in intercepts and slopes via random effects. However, multilevel models accomplish this by assuming that these intercepts and slopes come from a distribution of intercepts and slopes that are normally distributed. Thus, as opposed estimating 100 distinct intercept and 100 distinct slopes to capture between person heterogeneity in intercepts and slopes, multilevel models accomplish this by estimating two variances and a covariance.[2] The residual errors are also normally distributed, with unique residual variances for each outcome:

---

[2]Note that the when using maximum likelihood methods to estimate these models, the random effects are not estimated. Rather, the variance/covariance matrix of the random effects is estimated. The random effects can be predicted using empirical Bayes methods (Raudenbush & Bryk, 2002).

$$e_{1ij} \sim N\left(0, \sigma_{e1}^2\right), \quad (5)$$

$$e_{2ij} \sim N\left(0, \sigma_{e2}^2\right). \quad (6)$$

By fitting two independent models in Equations (1)–(6), we implicitly assume that depression and quality of life are independent. This is untenable as two outcomes from the same participant are almost certainly related. For example, change in depression is likely correlated with change in quality of life; thus, we should estimate the covariance between the random slopes (v_{1j} and v_{2j}). Further, because the outcomes are repeated measures on the same person, there will likely be a relationship between the residual errors (Fieuws & Verbeke, 2004). As discussed below, ignoring these relationships across outcomes will cause problems for tests of the difference between treatment effects.

A multivariate model can address these limitations because it allows us to model the relationships between the variables via correlations amongst the random effects and amongst the residuals (Hox, 2010; MacCallum et al., 1997). To help bridge the univariate and fully multivariate model, we describe a form of the multivariate multilevel model that makes the same assumptions as the two univariate models described above and is identical except that both outcomes are model simultaneously rather than sequentially. We then show how the assumptions inherent to the univariate models can be relaxed so that a fully multivariate model can be estimated.

In the univariate models, we described the random effects as coming from two distinct multivariate normal distributions described in Equations (3) and (4). In a multivariate model —that is, a model that includes both depression and quality of life simultaneously—the random effects are drawn from a single multivariate normal distribution (Snijders & Bosker, 2012). A multivariate normal distribution consistent with Equations (3) and (4) can be expressed as follows:

$$\begin{bmatrix} u_{1j} \\ v_{1j} \\ u_{2j} \\ v_{2j} \end{bmatrix} \sim MVN(0, \Omega_G), \Omega_G = \begin{bmatrix} \sigma_{u1}^2 & & & \\ \sigma_{u1v1} & \sigma_{v1}^2 & & \\ 0 & 0 & \sigma_{u2}^2 & \\ 0 & 0 & \sigma_{u2v2} & \sigma_{v2}^2 \end{bmatrix}. \quad (7)$$

The covariance matrix for the random effects in Equation (7), Ω_G, contains the same four variance components and the same two covariances as in Equations (3) and (4). However, Equation (7) makes explicit that we assume no relationship among the random effects across outcomes by constraining the between-outcome covariances to 0. For example, the covariance in the third row and first column is fixed to 0 and represents the covariance between the random intercept for depression and the random intercept for quality of life. Like the random effects, the multivariate model assumes the residuals come from a single multivariate normal distribution rather than two univariate normal distributions (Snijders & Bosker, 2012). A multivariate normal representation of Equations (5) and (6) is:

$$\left[\begin{array}{c} e_{1ij} \\ e_{2ij} \end{array} \right] \sim MVN(0, \Omega_R), \Omega_R = \left[\begin{array}{cc} \sigma_{e1}^2 & \\ 0 & \sigma_{e2}^2 \end{array} \right]. \quad (8)$$

We refer to a multivariate model the estimates the growth models in Equations (1) and (2) and covariance matrices described in Equations (7) and (8) as the *multivariate independent outcomes* model. The independent outcomes model is simply a multivariate version of our earlier univariate models that we use as a baseline to compare multivariate models that allow a relationship between the outcomes.

As noted previously, the problem with the independent outcomes model is that it assumes that depression and quality of life are unrelated. If we are interested in testing parameters that are fundamentally multivariate, the independence model will lead to problems. For example, if our goal was to understand whether CBT had a stronger effect relative to no treatment on depression than on quality of life (i.e., a parameter that involves multiple outcomes), then the independence model will produce an incorrect hypothesis test and confidence interval, as we show below.

A multivariate model that estimates the relationships between random effects across outcomes and residuals across outcomes will produce appropriate estimates and standard errors for multivariate parameters. We call this type of model a *multivariate related outcomes* model. In the related outcomes model, all covariances between each random effect are estimated:

$$\left[\begin{array}{c} u_{1j} \\ v_{1j} \\ u_{2j} \\ v_{2j} \end{array} \right] \sim MVN(0, \Omega_G), \ \Omega_G = \left[\begin{array}{cccc} \sigma_{u1}^2 & & & \\ \sigma_{u1v1} & \sigma_{v1}^2 & & \\ \sigma_{u1u2} & \sigma_{v1u2} & \sigma_{u2}^2 & \\ \sigma_{u1v2} & \sigma_{v1v2} & \sigma_{u2v2} & \sigma_{v2}^2 \end{array} \right]. \quad (9)$$

Thus, we now estimate the covariance between any combination of random intercepts and slopes—we no longer constrain the between-outcome covariances to 0 as we did in Equation (7). For example, the covariance between the random slopes, which describes the covariance between the rate of change in depression and quality of life, is σ_{v1v2}. The related outcomes model also estimates a covariance among the residual errors rather than constraining the covariance to 0 as we did in Equation (8):

$$\left[\begin{array}{c} e_{1ij} \\ e_{2ij} \end{array} \right] \sim MVN(0, \Omega_R), \Omega_R = \left[\begin{array}{cc} \sigma_{e1}^2 & \\ \sigma_{e1e2} & \sigma_{e2}^2 \end{array} \right]. \quad (10)$$

The only difference between the independent outcomes and related outcomes models are the covariances among the random effects across outcomes and the covariance among the residuals. These covariances can be important because they (a) can be interpreted substantively (e.g., what is the correlation between intercepts and slopes across the two outcomes?) and (b) impact statistical tests for multivariate parameters.

Data Set-up for Multivariate Models

Before turning to the application and results of the multivariate multilevel models, we briefly comment on data set-up. Multilevel modeling software developed from a structural equation model framework (e.g., Mplus) will typically expect the data to be in the wide format shown in Table 1. Consequently, multiple equations much like those represented above in Equations (1), (2), (9), and (10) can be specified. However, for most multilevel software packages (e.g., xtmixed in Stata, PROC MIXED in SAS, lme4 in R), we must combine Equations (1) and (2) into a single equation using a set of indicator variables to define which observations go with depression and quality of life. The long dataset in Table 1 provides an example of how the data must be organized for a single equation model. The values for depression and quality of life are combined into a single outcome variable, y_{hij}, where h indexes the outcome measure. We also create two indicator variables, d_j and q_j, where $d_j = 1$ for depression and 0 for quality of life and $q_j = 1$ for quality of life and 0 for depression. (Connecting the symbols to the actual column names in Table 1: $y_{hij} = Y$, $d_j =$ Depression, and $q_j =$ Quality of Life.) We then combine Equations (1) and (2) with the indicator variables to produce a single equation:

$$y_{hij}=\beta_{10}d_j+\beta_{20}q_j+\beta_{11}Time_{ij}d_j+\beta_{21}Time_{ij}q_j$$
$$+\beta_{12}Tx_jd_j+\beta_{22}Tx_jq_j+\beta_{13}Time_{ij}Tx_jd_j+\beta_{23}Time_{ij}Tx_jq_j \quad (11)$$
$$+u_{1j}d_j+u_{2j}q_j+v_{1j}Time_{ij}d_j+v_{2j}Time_{ij}q_j+e_{1ij}d_j+e_{2ij}q_j.$$

The random effects (u_{1j}, u_{2j}, v_{1j}, and v_{2j}) and residual errors (e_{1ij} and e_{2ij}) are distributed as in Equations (9) and (10), respectively.

We can verify that Equation (11) is identical to Equations (1) and (2) by examining the value of y_{hij} when plugging in the appropriate values of d_j and q_j for depression and quality of life. The value of y_{hij} for the depression outcome is:

$$(y_{hij}|d_j=1, q_j=0)=\beta_{10}+\beta_{11}Time_{ij}+\beta_{12}Tx_j+\beta_{13}Time_{ij}Tx_j+u_{1j}+v_{1j}Time_{ij}+e_{1ij}. \quad (12)$$

Likewise, the value of y_{hij} for quality of life is:

$$(y_{hij}|d_j=0, q_j=1)=\beta_{20}+\beta_{21}Time_{ij}+\beta_{22}Tx_j+\beta_{23}Time_{ij}Tx_j+u_{2j}+v_{2j}Time_{ij}+e_{2ij}. \quad (13)$$

Differential Treatment Effects

Using the simulated treatment data, we examined the differential treatment effects of CBT versus control on depression and quality of life. The online appendix material provides the data and annotated syntax for estimating these models in Stata, SPSS, SAS, Mplus, and R. We report the output from xtmixed in Stata using maximum likelihood estimation, although the software packages provide identical results out to 4 to 5 decimals places. In this section, we first compare the results and fit of the univariate, multivariate independent outcomes, and multivariate related outcomes models so that we can compare the results of each model. Second, we show how to test for differential treatment effects using a post analysis contrast and how the standard error for this contrast differs across the multivariate independent

outcomes and multivariate related outcomes models. Third, we describe two methods in addition to the post analysis contrast for testing differential treatment effects. Finally, testing differential treatment effects across outcomes requires that the outcomes be on the same metric. We discuss why outcomes need to be on the same metric and how one can use standardized scores in situations where outcomes are on different metrics and one wants to test differential effects.

Table 2 provides the estimates and standard errors for the fixed effects, the variance and covariance estimates for the random effects and residuals, and the deviance for each model. Table 2 makes clear that the independent outcomes model is identical to the two univariate models put together. The independent outcomes model estimates all the parameters in the two univariate models. Furthermore, the deviance, a measure of model fit (Singer & Willet, 2003), of the independent outcomes model is the sum of the deviance of the two univariate models. Thus, we can use the independent outcomes model as a comparison to investigate whether the multivariate related outcomes model improves upon the univariate models typically used in clinical research.

Because the independent outcomes model is nested within the related outcomes model, we can test whether the addition of the covariances among outcomes in the random effects and residuals significantly improves fit using a likelihood ratio test. The independent outcomes model is considered nested within the related outcomes model because the independent outcomes model (a) uses the same data as the related outcomes model and (b) is a constrained version of the related outcomes model (i.e., the between-outcome covariances are constrained to zero). The likelihood ratio test compares the difference between the deviances of the models (1501.63 − 1444.80 = 56.83; smaller is better) to a χ^2-distribution with degrees of freedom equal to the difference in the number of parameters between models (21 − 16 = 5). There are five additional parameters as the related outcomes model includes four additional covariances in the random effects of intercepts and slopes, and also allows the residual errors to be correlated. The likelihood ratio test equals $\chi^2(5)=56.83$, $p<.001$, which indicates that the related outcomes model fits the data better than the independent outcomes model. This test is also a joint test of the significance for the additional covariance parameters in the related outcomes model.

Although the related outcomes model has the best fit, Table 2 indicates that the parameter estimates do not change from independent model to fully multivariate model. Thus, it is reasonable to ask why should we bother with the multivariate model? All parameter estimates in Table 2, except for the covariances among the random effects between outcomes and covariance among residuals between outcomes, are univariate parameters— only data from one of the outcomes contributes to the univariate estimates. Consequently, the estimates and standard errors, for both the fixed effects and variance/covariances, are identical across models. In some cases, some parameters are not estimated and thus assumed zero but that does not affect the estimates of the univariate parameters. However, when we consider multivariate parameters, the differences between the independent outcomes and related outcomes models is important.

The additional covariances will affect tests of whether the treatment effects differ across outcomes. First, consider the univariate treatment effects for depression and quality of life, which are the two time×treatment interactions terms, β_{13} and β_{23}. Each coefficient describes the difference in rate of change among the intervention conditions for each outcome separately, with the null hypothesis being that there is no difference between conditions. The time×treatment interaction for depression was $\beta_{13}=-.41$ ($p < .05$, 95% CI = $-.60, -.22$) whereas the time×treatment interaction for quality of life was $\beta_{23}=-.09$ ($p = .41$, 95% CI = $-.30, .12$). Thus, we reject the null hypothesis of no treatment effect for depression but not for quality of life. We may be tempted to also conclude that the effects of treatment versus control are larger for depression than quality of life. However, this would be incorrect because we have only tested the univariate null hypothesis that each coefficient is zero. We have not tested whether the coefficients differ from one another—that is, we have not tested the multivariate null hypothesis that $\beta_{13} - \beta_{23}=0$.

To test the null hypothesis that $\beta_{13} - \beta_{23}=0$ we can use a post-analysis contrast. A common form for this test is:

$$z=\frac{\beta_{13}-\beta_{23}}{se_{\beta_{13}-\beta_{23}}}; se_{\beta_{13}-\beta_{23}}=\sqrt{\sigma^2_{\beta_{13}}+\sigma^2_{\beta_{23}}-2\sigma_{\beta_{13}\beta_{23}}}, \quad (14)$$

where the denominator is the standard error of the difference between treatment effects, $\sigma^2_{\beta_{13}}$ and $\sigma^2_{\beta_{23}}$ are the expected variability of the treatment effect across samples and $\sigma_{\beta_{13}\beta_{23}}$ is the covariance between the treatment effects across samples.[3] As we have seen, the independent outcomes and related outcomes models will provide identical estimates of β_{13} and β_{23}, so the numerator in Equation (14) will be identical across models. Thus, the key part of Equation (14) is the standard error, specifically the covariance, $\sigma_{\beta_{13}\beta_{23}}$. If depression and quality of life are correlated, then independent outcomes misspecifies the relationships among the random effects and among the residuals and $\sigma_{\beta_{13}\beta_{23}}$ will incorrectly be set to 0. In contrast, the related outcomes model correctly specifies the relationships among the random effects and residuals and $\sigma_{\beta_{13}\beta_{23}}$ will be positive. The misspecification in the independent outcomes model means the standard error will be too large and will thus reduce power.

The test of differential treatment effects was significant in both the independent outcomes ($\beta_{13} - \beta_{23}=-0.32$, $se = 0.15$, $z = -2.23$, $p = 0.03$) and related outcomes model ($\beta_{13} - \beta_{23}= -0.32$, $se = 0.10$, $z = -3.14$, $p = 0.002$). Note, however, that the standard error in the independent outcomes model was 50% larger than in the related outcomes model because the independent outcomes model does not take into account the correlation among the outcomes. To provide a sense of how the larger standard error will impact power, we simulated 10,000 additional datasets using the population parameters described above and fit the independent and related outcomes models to each dataset. To assess power, we computed how often the test of the differential treatment effects was significant across the 10,000 datasets. Power was 0.81 for the related outcomes model and 0.68 for the independent outcomes model, a 16% decrease in power.

---

[3]These variances and covariances are part of the asymptotic variance/covariance matrix of the fixed effects. See the online appendix for instructions how to request this information within statistical packages.

Equation (14) is not the only method for evaluating differential intervention effects. A second method is to use a likelihood ratio test to compare the fit of a model that estimates distinct treatment effects by outcome such as Equation (11) to a model that estimates a common treatment effect for outcomes. The model with a common treatment effect is identical to Equation (11) except in the second model we estimate a common time×treatment interaction (underlined):

$$
\begin{aligned}
y_{hij} = {} & \beta_{10}d_j + \beta_{20}q_j + \beta_{11}\,Time_{ij}d_j + \beta_{21}\,Time_{ij}q_j \\
& + \beta_{12}Tx_jd_j + \beta_{22}Tx_jq_j + \underline{\beta_3\,Time_{ij}Tx_j} \\
& + u_{1j}d_j + u_{2j}q_j + v_{1j}\,Time_{ij}d_j + v_{2j}\,Time_{ij}q_j + e_{1ij}d_j + e_{2ij}q_j.
\end{aligned}
\tag{15}
$$

The common time × treatment is constructed by multiplying the time variable and the treatment indicator irrespective of outcome type. This is in contrast to Equation (11) where the time × treatment interaction is multiplied by the outcome variable indicators so that unique treatment effects can be estimate for each outcome. The null hypothesis for the likelihood ratio test is that there is no difference in model fit between the model with a common time×treatment interaction—Equation (15)—and the model with unique time×treatment interactions—Equation (11).[4] In our example, the likelihood ratio test was significant, $\chi^2(1)=9.43$, $p<.01$, indicating that the model with unique treatment effects for outcome types fit the data better than the model with a common treatment effect.

The third method for testing for differential treatment effects is to add a three-way interaction between time, treatment, and either d_j or q_j to Equation (15):

$$
\begin{aligned}
y_{hij} = {} & \beta_{10}d_j + \beta_{20}q_j + \beta_{11}\,Time_{ij}d_j + \beta_{21}\,Time_{ij}q_j \\
& + \beta_{12}Tx_jd_j + \beta_{22}Tx_jq_j + \underline{\beta_3\,Time_{ij}Tx_j} + \underline{\beta_4\,Time_{ij}Tx_jd_j} \\
& + u_{1j}d_j + u_{2j}q_j + v_{1j}\,Time_{ij}d_j + v_{2j}\,Time_{ij}q_j + e_{1ij}d_j + e_{2ij}q_j.
\end{aligned}
\tag{16}
$$

It does not matter whether we use d_j or q_j; however, because we used d_j then β_3 is interpreted as the time×treatment interaction for quality of life and β_4 is difference between the time×treatment interaction for depression and the time×treatment interaction for quality of life. Thus, the null hypothesis for β_4 is that there is no difference between the time×treatment interaction across outcomes—that is, the treatment effect does not differ with respect to outcome. The significance test for β_4 is identical to the post-analysis contrast of the difference between β_{13} and β_{23} from Equation (11).

Suppose that none of the tests of the differential treatment effect was significant. This would indicate that the outcomes share a common treatment effect. In that case, the model in Equation (15), which estimates a common treatment effect across outcomes, could be used in place of the model in Equation (11), which estimates unique treatment effects. Note that this simplification is justified because we formally tested whether the coefficients differed.

---

[4]Because Equations (11) and (15) differ with respect to the fixed effects, the likelihood ratio test can only be used if maximum likelihood rather than restricted maximum likelihood estimation was used (see Singer & Willet, 2003, for an introductory disscussion of likelihood ratio tests).

Simply evaluating the significance of the coefficients in Equation (11) does not provide a basis for estimating a common effect.

Although we have focused on using these tests to evaluate differential treatment effects across outcomes, the three methods we described can be used to test whether any fixed effect differs across outcomes. For example, in a psychopathology study, we could use these methods to evaluate whether the average growth in one variable differs from the average growth in a second variable. Furthermore, these tests can be extended to more than two outcomes in which once can test whether all effects are equal.

In order for the tests of differential treatment effects, or any test of differences between fixed effects, to be interpretable, it is essential that the outcome variables be on the same metric as they were in our example. Recall that the null hypothesis for the differential treatment effect is that $\beta_{13} - \beta_{23} = 0$. The scale of β_{13} and β_{23} are dictated by the scale of the outcome variable. If depression and quality of life are on different metrics then the difference between the coefficients may be non-zero as function of the metrics rather than a real difference. Consequently, the difference will be difficult, if not impossible, to interpret and the hypothesis test incorrect. This is precisely the same problem faced in meta-analysis. In a meta-analysis of randomized trials we aim to compare treatment effects across different measures with different scales. To solve this problem we standardize the treatment effects using effect sizes. In treatment meta-analyses, we often use Cohen's *d*, which expresses mean differences in terms of standard deviation units. We recommend a similar strategy for multivariate models where the outcome variables are on different metrics and where the aim of the analysis is to test differences in fixed effects across outcomes. The standardization could take several forms but a reasonable choice would be to create *z*-scores using the mean and standard deviation from all time points.

Examining the Relationship Between Outcomes

Up to this point we have used the multivariate multilevel model to examine whether the relationship between a predictor variable (e.g., treatment condition) and the outcome variable differed across outcomes type (e.g., depression versus quality of life)—that is, we compared the fixed effects across outcomes. We can also use the multivariate model to examine the relationship between the outcomes. Suvak et al. (2009) used multivariate multilevel models to explore the relationship between intrusion and avoidance over time in a sample of trauma survivors. For example, they examined the correlation between rate of change in intrusion symptoms and avoidance symptoms (i.e., the correlation between the random slopes) as well as the correlation between initial intrusion symptoms and rate of change in avoidance symptoms (i.e., the correlation between the random intercept and random slope) and vice versa.

In our example data, we considered the two categories of correlations used by Suvak et al. (2009). The first category of correlation is to examine the relationship between similar parameters across outcomes. We can examine the relationship between the random intercepts from each outcome to assess how the person-specific baseline values are related. Likewise, we can examine the relationship between the random slopes from each outcome to

assess how the person-specific rates of change are related. The formulae for the correlation between intercepts, r_I, and slopes, r_S, are (Fieuws & Verbeke, 2004):

$$r_I = \frac{\sigma_{u_1 u_2}}{\sqrt{\sigma_{u_1}^2}\sqrt{\sigma_{u_2}^2}} \quad (17)$$

$$r_S = \frac{\sigma_{v_1 v_2}}{\sqrt{\sigma_{v_1}^2}\sqrt{\sigma_{v_2}^2}} \quad (18)$$

Each of the components in Equations (17) and (18) are drawn from the covariance matrix of the multivariate related outcomes model described in Equation (9) and displayed in Table 2. Thus, the numerator of Equation (17) is the covariance between the random intercepts and the denominator is the product of the square root of the variance of the intercepts. Equation (18) is identical except it involves the variances and covariance of the random slopes. The correlation among the intercepts was $r_I = .26$,[5] indicating a small positive relationship between the person-specific baseline values of depression and quality of life. The correlation among the slopes was $r_S = .76$, indicating a strong positive relationship between the person-specific rates of change for depression and quality of life.

Power can be a significant challenge when estimating the relationship between slopes. Hertzog, Lindenberger, Ghisletta, and von Oertzen (2006) showed that power to detect relationships among slopes can be low even with large samples and several measurement occasions.[6] A key factor in determining power is what they called Growth Curve Reliability, which can be defined as the proportion of the total variability in an outcome that is accounted for by person-to-person variability in intercepts and slopes. If the reliability is low, power for examining the relationship between slopes will typically be low, even when sample sizes are large. Researchers wishing to examine relationships between outcomes will need to consider these issues when designing their studies and interpreting the results of the multivariate models.

The second category of correlation is to examine the relationship between distinct parameters across outcomes. Substantively, we might consider whether baseline quality of life affects the rate of change in depression symptoms. The formula for the correlation between the random intercepts for quality of life and the rate of change for depression is:

$$r_{IS} = \frac{\sigma_{u2v1}}{\sqrt{\sigma_{u2}^2}\sqrt{\sigma_{v1}^2}}. \quad (19)$$

---

[5]Standard errors for this correlation, as well as the others described in this section, can be obtained using the delta method (Fieuws & Verbeke, 2004). Stata and Mplus will provide standard errors based on the delta method. However, the delta method assumes the sampling distribution of these correlations is normally distributed, which it is not. Consequently, the delta method can produce problematic results, such as confidence intervals that exceed the boundary of a correlation. Alternatives include fitting a model with a covariance a model with the covariance constrained to zero and using a likelihood ratio test to compare the fit. Additionally, bootstrapping the confidence intervals can be helpful. Finally, Bayesian methods provide a useful alternative for interval estimation for correlations (Baldwin & Fellingham, 2012).
[6]We thank an anonymous reviewer for pointing us toward this reference.

The formula the correlation between the random intercepts for depression and the rate of change for quality of life has the same form but uses the appropriate parameters. There was a small negative correlation between baseline values for quality of life and rate of change in depression, $r_{IS} = -.23$. The correlation between baseline values in depression and rate of change in quality of life was also small and negative, $r_{IS} = -.16$. Thus, there was little relationship between a participant's baseline standing on one outcome and the rate of change on the other outcome.

Extending the Model to Three or More Outcomes

The multivariate model can be extended to three or more outcomes by adding fixed and random effects for the third outcome. For example, suppose in addition to depression and quality of life we also measured anxiety symptoms. To Equations (1) and (2) we would add the following equation:

$$y_{3ij} = \beta_{30} + \beta_{31} Time_{ij} + \beta_{32} Tx_j + \beta_{33} Time_{ij} Tx_j + u_{3j} + v_{3j} Time_{ij} + e_{3ij}, \quad (20)$$

where the parameters have the same interpretation as before except the pertain to anxiety. In order to accommodate the model with three outcomes in software that expects the data in long format, we would again use indicator variables to identify which parameters go with particular outcomes. In this case, we create three indicator variables, d_j, q_j, and a_j, where $d_j = 1$ for depression and 0 for the others, $q_j = 1$ for quality of life and 0 for the others, $a_j = 1$ for anxiety and 0 for the others. As previously, we use the indicator variables to create a single equation:

$$\begin{aligned} y_{hij} = {} & \beta_{10} d_j + \beta_{20} q_j + \beta_{30} a_j + \beta_{11} Time_{ij} d_j + \beta_{21} Time_{ij} q_j + \beta_{31} Time_{ij} a_j \\ & + \beta_{12} Tx_j d_j + \beta_{22} Tx_j q_j + \beta_{32} Tx_j a_j + \beta_{13} Time_{ij} Tx_j d_j + \beta_{23} Time_{ij} Tx_j q_j + \beta_{33} Time_{ij} Tx_j a_j \\ & + u_{1j} d_j + u_{2j} q_j + u_{3j} a_j + v_{1j} Time_{ij} d_j + v_{2j} Time_{ij} q_j + v_{3j} Time_{ij} a_j \\ & + e_{1ij} d_j + e_{2ij} q_j + e_{3ij} a_j \end{aligned} \quad (21)$$

The variance/covariance matrices for the random effects and the residuals are also extended. The random effects matrix includes all six random effects:

$$\begin{bmatrix} u_{1j} \\ v_{1j} \\ u_{2j} \\ v_{2j} \\ u_{3j} \\ v_{3j} \end{bmatrix} \sim MVN(0, \Omega_G), \Omega_G = \begin{bmatrix} \sigma_{u1}^2 & & & & & \\ \sigma_{u1v1} & \sigma_{v1}^2 & & & & \\ \sigma_{u1u2} & \sigma_{v1u2} & \sigma_{u2}^2 & & & \\ \sigma_{u1v2} & \sigma_{v1v2} & \sigma_{u2v2} & \sigma_{v2}^2 & & \\ \sigma_{u1u3} & \sigma_{v1u3} & \sigma_{u1u3} & \sigma_{v2u3} & \sigma_{u3}^2 & \\ \sigma_{u1v3} & \sigma_{v1v3} & \sigma_{u2v3} & \sigma_{v2v3} & \sigma_{u3v3} & \sigma_{v3}^2 \end{bmatrix}. \quad (22)$$

The residual matrix includes all three residuals:

$$\begin{bmatrix} e_{1ij} \\ e_{2ij} \\ e_{3ij} \end{bmatrix} \sim MVN(0, \Omega_R), \Omega_R = \begin{bmatrix} \sigma_{e1}^2 & & \\ \sigma_{e1e2} & \sigma_{e2}^2 & \\ \sigma_{e1e3} & \sigma_{e2e3} & \sigma_{e3}^2 \end{bmatrix}. \quad (23)$$

For both matrices the interpretation of the parameters is the same as discussed previously except now the matrices include additional variances and covariances involving the third outcome.

The types of question we can examine with three or more outcomes are similar to the model with two outcomes. We can compute the correlations among the random effects. Likewise, we can compare whether treatment effects differ across the three outcomes or whether the treatment effect on depression differs from the average of the other outcomes. As before, scaling of the outcome variables needs to be identical for the comparisons to be meaningful. We can also estimate a joint test of whether all three treatment effects are zero using a post-analysis contrast. However, the simplest way is to use a likelihood ratio test that compares the fit of the model in Equation (21) to a model that is identical to Equation (21) except that it excludes all time × treatment effects (i.e., fixes all treatment effects to zero).

The model with three outcomes is considerably more complex than even the model with two outcomes. For example, the model with three outcomes involves 39 parameters and the model with two outcomes involves 21 parameters. It can be challenging to ensure that the more complicated model is correctly specified in software programs as there are many variables and interactions in both the fixed and random effects portions of the model. Furthermore, most multilevel software was not explicitly designed to model multivariate data in this way. Consequently, estimation can be slow, especially with large datasets. Multilevel software that is multivariate in nature, such as Mplus, typically does not have much trouble estimating these models, even with large datasets.

Extensions

In the present paper, we detailed how to use multivariate multilevel models to examine the relationship between multiple outcomes in a clinical trial. However, this is one of many possible applications of multivariate models that psychotherapy researchers might use. For example, models for dyadic psychotherapy data can be framed as a multivariate model. Suppose one had collected martial satisfaction data three times for 100 couples. One way to model this data is to use a 3-level multilevel model, with observations at level-1, participant at level-2, and dyad at level-3. Alternatively, we could use a 2-level multivariate model with two outcomes, one for each partner (Atkins, 2005, describes the benefits and drawbacks for both 2- and 3-level models). This multivariate model is similar to our differential treatment effects examples, except we would exchange the primary and secondary outcomes with the partners' outcomes. The actor-partner interdependence model for dyadic data can also be estimated using a multivariate multilevel model (Kenny, Kashy, & Cook, 2006).

Multivariate multilevel models can be used in social relations modeling (Kenny, 1994; Kenny et al., 2006) approaches to psychotherapy data. For example, Marcus, Kashy, and Baldwin (2009) used a multivariate multilevel model to apply a social relations model to therapeutic alliance data (see also Marcus, Kashy, Wintersteen, & Diamond, 2011). In this study, patients and therapists each rated the alliance. The primary question was whether therapists and clients agree with respect to their alliance ratings. That is, if a therapist consistently rates her alliances as high across her caseload, do her patients also rate the

alliance high across her caseload—the correlation among the patient and therapist alliance ratings at the therapist-level. Additionally, they examined whether therapists and patients agree within caseloads—the correlation among the patient and therapist ratings at the patient-level (see also Imel, Hubbard, Rutter, & Simon, 2013).

Other examples of multivariate multilevel models relevant to psychotherapy data include using them: (a) to examine mediation hypotheses in a multilevel context (Bauer, Preacher, & Gil, 2006), in which both the dependent variable and the mediator(s) are included as outcomes in the multilevel model; (b) to fit measurement models where item responses are included as outcomes in the multilevel model (Zheng & Rabe-Hesketh, 2007); (c) to fit multivariate meta-analyses where two or more effect sizes from a single study are analyzed simultaneously (Riley, Abrams, Sutton, Lambert, & Thompson, 2007; White, 2011); (d) to examine the relationship between outcomes with different distributional forms (e.g., comparing a normally distributed outcome to a binary outcome; Hadfield, 2010); and (e) to simultaneously examine predictors unique to each outcome or different functional forms for change across outcomes (e.g., linear change for one outcome and quadratic change for another; MacCallum et al., 1997).

Another extension to the models we have presented is to estimate them from a structural equation modeling perspective or from a perspective that combines multilevel modeling and structural equation modeling.[7] For example, Bollen and Curran (2006) discuss a multivariate latent growth curve model similar to our example but where one estimates regression paths between the random intercepts and slopes, rather than covariances only. For example, in an alcohol treatment study, if we hypothesized that variability among patients in depressive symptoms prior to treatment (i.e., random intercept for depression) would predict variability in the rate of change in drinking behavior (i.e., random slope for drinking behavior), we could estimate the regression path between the random intercept for depression and random slope for drinking behavior.

Mehta and Neale (2005) also illustrate how multilevel factor analysis is an extension of the multivariate multilevel models. They fit a multilevel factor analysis model to literacy data that included five measures of literacy taken on students clustered within schools. They fit a multilevel model with random intercepts for each of the schools. Rather than estimating correlations among the random intercepts and residuals, they estimated a common factor at both the school-level (i.e., random intercept) and student-level (i.e., residual). To be sure, fitting such a model is challenging and may require more data than is typically available to psychotherapy researchers. Nevertheless, Mehta and Neale's (2005) example provides an additional illustration of how we could extend the models discussed in this paper. Other interesting methodological developments are likely to occur at the intersection of multilevel modeling and structural equation modeling.

---

[7]The intersection between structural equation modeling and multilevel models has been discussed at length (e.g., Bauer, 2003; Curran, 2003; Skrondal & Rabe-Hesketh, 2004), with a number of writers illustrating how one can parameterize structural equation models to reproduce a multilevel model (Bauer, 2003; Curran, 2003; Mehta & Neale, 2005). Theoretical work has been done to frame multilevel models and structural equation models as special cases of a broader class of models known as *generalized latent variable models* (Skrondal & Rabe-Hesketh, 2004). In fact, the notion that multilevel models and structural equation models are both latent variable models is a foundational idea for some software programs including gllamm (which stands for Generalized Linear Latent and Mixed Models) in Stata (Rabe-Hesketh, Skrondal, & Pickles, 2002) and Mplus (Muthén & Muthén, 2012).

Conclusion

We introduced the multivariate multilevel model as a way to examine hypotheses that are important to psychotherapy research. As with any methodology, advances in psychotherapy research are not going to take place simply because we apply multivariate multilevel models to our data. However, many, if not most, theories about how and why psychotherapy works are multivariate in nature. As we have shown, these multivariate hypotheses can be addressed with a multivariate multilevel model in ways that are not possible with univariate models. We suspect as more researchers become familiar with these techniques we will see more creative uses of multivariate models that improve our understanding of psychotherapy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Atkins DC. Using multilevel models to analyze couple and family treatment data: Basic and advanced issues. Journal of Family Psychology. 2005; 19:98–110. [PubMed: 15796656]

Baldwin SA, Bauer DJ, Stice E, Rohde P. Evaluating models for partially clustered designs. Psychological Methods. 2011; 16:149–165. [PubMed: 21517179]

Baldwin SA, Fellingham GW. Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. Psychological Methods. 2012 Advance Online Publication. 10.1037/a0030642

Baldwin, SA.; Imel, ZE. Therapist effects: Findings and methods. In: Lambert, MJ., editor. Bergin and Garfield's Handbook of Psychotherapy and Behavior Change. 6. New York: Wiley; 2013. p. 258-297.

Baldwin SA, Murray DM, Shadish WR. Empirically supported treatments or Type I errors? Problems with the analysis of data from group-administered treatments. Journal of Consulting and Clinical Psychology. 2005; 73:924–935. [PubMed: 16287392]

Bauer DJ. Estimating multilevel linear models as structural equation models. Journal of Educational and Behavioral Statistics. 2003; 28:135–167.

Bauer DJ, Preacher KJ, Gil KM. Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. Psychological Methods. 2006; 11:142–163. [PubMed: 16784335]

Bauer DJ, Sterba SK, Hallfors DD. Evaluating group-based interventions when control participants are ungrouped. Multivariate Behavioral Research. 2008; 43:210–236. [PubMed: 20396621]

Bollen, KA.; Curran, PJ. Latent curve models: A structural equation perspective. Hoboken, NJ: Wiley; 2006.

Crits-Christoph P, Baranackie K, Kurcias JS, Beck AT, Carroll K, Perry K, Zitrin C. Meta-analysis of therapist effects in psychotherapy outcome studies. Psychotherapy Research. 1991; 1:81–91.

Crits-Christoph P, Mintz J. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. Journal of Consulting and Clinical Psychology. 1991; 59:20–26. [PubMed: 2002139]

Curran PJ. Have multilevel models been structural equation models all along? Multivariate Behavioral Research. 2003; 38:529–569.

Fieuws S, Verbeke G. Joint modelling of multivariate longitudinal profiles: Pitfalls of the random-effects approach. Statistics in Medicine. 2004; 23:3093–3104.10.1002/sim.1885 [PubMed: 15449333]

Gelman, A.; Hill, J. Data analysis using regression and multilevel/hierarchical models. New York: Cambridge University Press; 2007.

Hadfield JD. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. Journal of Statistical Software. 2010; 33:1–22. [PubMed: 20808728]

Hertzog C, Lindenberger U, Ghisletta P, Oertzen Tv. On the power of multivariate latent growth curve models to detect correlated change. Psychological Methods. 2006; 11:244–252.10.1037/1082-989X.11.3.244 [PubMed: 16953703]

Hofmann SG, Lohr JM. To kill a dodo bird. The Behavior Therapist. 2010; 33:14–15.

Hox, JJ. Multilevel analysis: Techniques and applications. 2. New York: Routledge; 2010.

Imel ZE, Baldwin SA, Bonus K, Maccoon D. Beyond the individual: Group effects in mindfulness-based stress reduction. Psychotherapy Research. 2008; 18:735–742. [PubMed: 18815948]

Imel ZE, Hubbard RA, Rutter CM, Simon G. Patient-rated alliance as a measure of therapist performance in two clinical settings. Journal of Consulting and Clinical Psychology. 2013; 81:154–165.10.1037/a0030903.supp [PubMed: 23231576]

Jouriles EN, McDonald R, Rosenfield D, Stephens N, Corbitt-Shindler D, Miller PC. Reducing conduct problems among children exposed to intimate partner violence: A randomized clinical trial examining effects of project support. Journal of Consulting and Clinical Psychology. 2009; 77:705–717.10.1037/a0015994 [PubMed: 19634963]

Kaysen D, Atkins DC, Moore SA, Lindgren KP, Dillworth T, Simpson T. Alcohol Use, Problems, and the Course of Posttraumatic Stress Disorder: A Prospective Study of Female Crime Victims. Journal of Dual Diagnosis. 2011; 7:262–279.10.1080/15504263.2011.620449 [PubMed: 23538605]

Kenny, DA. Interpersonal perception: A social relations analysis. New York: Guilford; 1994.

Kenny, DA.; Kashy, DA.; Cook, WL. Analysis of dyadic data. New York: Guilford; 2006.

MacCallum RC, Kim C, Malarkey WB, Kiecolt-Glaser JK. Studying multivariate change using multilevel models and latent curve models. Multivariate Behavioral Research. 1997; 32:215–253.10.1207/s15327906mbr3203_1

Marcus DK, Kashy DA, Baldwin SA. Studying psychotherapy using the one-with-many design: The therapeutic alliance as an exemplar. Journal of Counseling Psychology. 2009; 56:537–548.10.1037/a0017291

Marcus DK, Kashy DA, Wintersteen MB, Diamond GS. The therapeutic alliance in adolescent substance abuse treatment: A one-with-many analysis. Journal of Counseling Psychology. 2011; 58:449–455.10.1037/a0023196 [PubMed: 21517154]

McDonagh A, Friedman M, McHugo G, Ford J, Sengupta A, Mueser K, Descamps M. Randomized trial of cognitive-behavioral therapy for chronic posttraumatic stress disorder in adult female survivors of childhood sexual abuse. Journal of Consulting and Clinical Psychology. 2005; 73:515–524.10.1037/0022-006X.73.3.515 [PubMed: 15982149]

Mehta PD, Neale MC. People are variables too: Multilevel structural equations modeling. Psychological Methods. 2005; 10:259–284.10.1037/1082-989X.10.3.259 [PubMed: 16221028]

Muthén, LK.; Muthén, BO. Mplus user's guide. 7. Los Angeles, CA: Muthén & Muthén; 2012.

Nieuwenhuis S, Forstmann BU, Wagenmakers EJ. Erroneous analyses of interactions in neuroscience: A problem of significance. Nature neuroscience. 2011; 14:1105–1107.10.1038/nn.2886

Rabe-Hesketh, S.; Skrondal, A. Multilevel and longitudinal modeling using Stata. College Station, TX: Stata Press; 2008.

Rabe-Hesketh S, Skrondal A, Pickles A. Reliable estimation of generalized linear mixed models using adaptive quadrature. The Stata Journal. 2002; 2:1–21.

Raudenbush, SW.; Bryk, AS. Hierarchical linear models: Applications and data analysis methods. Thousand Oaks, CA: Springer; 2002.

Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. BMC Medical Research Methodology. 2007; 7:3.10.1186/1471-2288-7-3 [PubMed: 17222330]

Saxon D, Barkham M. Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. 2012; 80:535–546.10.1037/a0028898

Singer, JD.; Willet, JB. Applied longitudinal data analysis: Modeling change and event occurence. New York: Oxford University Press; 2003.

Skrondal, A.; Rabe-Hesketh, S. Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. Boca Raton, FL: Chapman & Hall/CRC Press; 2004.

Snijders, TAB.; Bosker, RJ. Multilevel analysis: An introduction to basic and advanced multilevel modeling. 2. Thousand Oaks, CA: Sage; 2012.

Suvak MK, Walling SM, Iverson KM, Taft CT, Resick PA. Multilevel regression analyses to investigate the relationship between two variables over time: Examining the longitudinal association between intrusion and avoidance. Journal of traumatic stress. 2009; 22:622–631.10.1002/jts.20476 [PubMed: 19937725]

Wampold BE, Serlin RC. The consequences of ignoring a nested factor on measures of effect size in analysis of variance. Psychological Methods. 2000; 5:425–433. [PubMed: 11194206]

White IR. Multivariate random-effects meta-regression: Updates to mvmeta. The Stata Journal. 2011; 11:255–270.

Zheng X, Rabe-Hesketh S. Estimating parameters of dichotomous and ordinal item response models with gllamm. The Stata Journal. 2007; 7:313–333.

Table 1

Wide and Long Datasets for Multivariate Multilevel Data.

Wide			
ID	Time	Depression	Quality of Life
1	0	−1.37	−0.58
1	1	−1.31	0.24
1	2	−0.52	0.01
2	0	0.05	2.00
2	1	−0.84	1.83
2	2	−1.86	0.31

Long				
ID	Time	Y	Depression	Quality of Life
1	0	−1.37	1	0
1	1	−1.31	1	0
1	2	−0.52	1	0
1	0	−0.58	0	1
1	1	0.24	0	1
1	2	0.01	0	1
2	0	0.05	1	0
2	1	−0.84	1	0
2	2	−1.86	1	0
2	0	2.00	0	1
2	1	1.83	0	1
2	2	0.31	0	1

Table 2

Results of Univariate and Multivariate Models for Depression and Quality of Life Outcomes.

| | | Univariate | | Multivariate | |
	Parameter	Depression	Quality of Life	Independent Outcomes	Related Outcomes
Depression					
Intercept	β_{10}	−.08 (.13)	--	−.08 (.13)	−.08 (.13)
Time	β_{11}	.06 (.07)	--	.06 (.07)	.06 (.07)
Treatment	β_{12}	.16 (.18)	--	.16 (.18)	.16 (.18)
Time × Treatment	β_{13}	−.41* (.10)	--	−.41* (.10)	−.41* (.10)
Quality of Life					
Intercept	β_{20}	--	−.08 (.14)	−.08 (.14)	−.08 (.14)
Time	β_{21}	--	.002 (.08)	.002 (.08)	.002 (.08)
Treatment	β_{22}	--	.08 (.20)	.08 (.20)	.08 (.20)
Time × Treatment	β_{23}	--	−.09 (.14)	−.09 (.14)	−.09 (.14)
Depression					
u_{1j}	σ^2_{u1}	.49	--	.49	.49
v_{1j}	σ^2_{v1}	.04	--	.04	.04
cov(u_{1j},v_{1j})	σ_{u1v1}	.004	--	.004	.004
Quality of Life					
u_{2j}	σ^2_{u2}	--	.73	.73	.73
v_{2j}	σ^2_{v2}	--	.14	.14	.14
cov(u_{2j},v_{2j})	σ_{u2v2}	--	−.14	−.14	−.14
Between Outcomes					
cov(u_{1j},u_{2j})	σ_{u1u2}	--	--	--	.16
cov(u_{1j},v_{2j})	σ_{u1v2}	--	--	--	−.04
cov(v_{1j},u_{2j})	σ_{v1u2}	--	--	--	−.04
cov(v_{1j},v_{2j})	σ_{v1v2}	--	--	--	.06

| | Univariate | | Multivariate | |
Parameter	Depression	Quality of Life	Independent Outcomes	Related Outcomes
Residuals				
e_{1ij} σ^2_{e1}	.39	--	.39	.39
e_{2ij} σ^2_{e2}	--	.30	.30	.30
$\text{cov}(e_{1ij}e_{2ij})$ σ_{e1e2}	--	--	--	.15
Deviance	752.268	749.363	1501.631	1444.798

Note.

*
p < .05