Organizational
Research Methods
Volume 12 Number 4
October 2009 695-719

10.1177/1094428108327450
http://orm.sagepub.com
hosted at
http://online.sagepub.com

Testing Multilevel Mediation Using Hierarchical Linear Models

Problems and Solutions

Zhen Zhang
Arizona State University
Michael J. Zyphur
University of Washington, Bothell
Kristopher J. Preacher
University of Kansas

Testing multilevel mediation using hierarchical linear modeling (HLM) has gained tremendous popularity in recent years. However, potential confounding in multilevel mediation effect estimates can arise in these models when within-group effects differ from between-group effects. This study summarizes three types of HLM-based multilevel mediation models, and then explains that in two types of these models confounding can be produced and erroneous conclusions may be derived when using popularly recommended procedures. A Monte Carlo simulation study illustrates that these procedures can underestimate or overestimate true mediation effects. Recommendations are provided for appropriately testing multilevel mediation and for differentiating within-group versus between-group effects in multilevel settings.

Keywords: *multilevel; mediation; hierarchical linear models; random coefficient regression*

Using hierarchical linear modeling (HLM) to examine multilevel relationships has become a popular research practice among scholars in the last two decades (Klein & Kozlowski, 2000). Researchers are often interested in the top-down influences of higher-level constructs on lower-level constructs or relationships (Hofmann, Griffin, & Gavin, 2000; Snijders & Bosker, 1999). One benefit of multilevel models is that they allow researchers to hypothesize and empirically test interesting questions about multilevel mediation processes that are not easily answered using conventional statistical procedures (Mathieu, DeShon, & Bergh, 2008). Although the analytical challenges for mediation testing in multilevel settings have been discussed (e.g., MacKinnon, 2008; Mathieu et al. 2008; Mathieu & Taylor, 2007), an explicit examination of these challenges, the severity of the problems, and detailed recommendations for addressing them have not yet been offered.

In this study, we provide an in-depth exploration of one important issue to consider when conducting tests of multilevel mediation. Specifically, we focus on the confounding of within-group and between-group effects when conducting these tests. Although many

Authors' Note: An earlier version of this article was presented at the 2008 Annual Meeting of the Academy of Management held in Anaheim, California. Part of this research was conducted while the second author was a faculty member at the National University of Singapore. We are grateful to the editor Robert Vandenberg and three anonymous reviewers for their constructive comments.

researchers have explored the topic of differentiating effects across levels of analysis (e.g., Dansereau, Alutto, & Yammarino, 1984; Klein & Kozlowski, 2000), these treatments do not cover the confounding of effects in the context of multilevel mediation.

First, we provide a brief summary of the three types of multilevel mediation models that have been extensively used by researchers—these models have been discussed, for example, by Krull and MacKinnon (2001) and Mathieu and Taylor (2007). Then, effects that function within-groups versus between-groups are differentiated and concrete examples of such effects are provided. We show how potential confounding in mediation effects can arise and incorrect substantive conclusions can be drawn when researchers are not sensitive to these effects at different levels. The available guidelines for testing multilevel mediation (e.g., Krull & MacKinnon, 2001) are based on straightforward reformulations of single-level techniques (e.g., Baron & Kenny, 1986), and these procedures may produce underestimates or overestimates of the true multilevel mediation effects of interest. This is because typical HLM models often conflate between-group and within-group effects, where these effects at different levels should be examined separately.

This study contributes to the research methods literature by explicitly showing the consequences of conflating effects at different levels of analysis and by offering recommendations to address this issue. Based on a simulation study, we show that when a within-group effect is larger or smaller than a between-group effect, the popularly recommended HLM procedures could produce confounded estimates that are larger or smaller, respectively, than the true mediation effect. We show that the commonly used Sobel test (Sobel, 1982) and Freedman and Schatzkin's (1992) test of the mediation effect can be erroneous when the within- and the between-group effects differ in magnitude. In addition, very high Type-I error rates can be found when mediation effects are overestimated. In our discussion, we put forth a set of recommendations for appropriately conducting multilevel mediation testing.

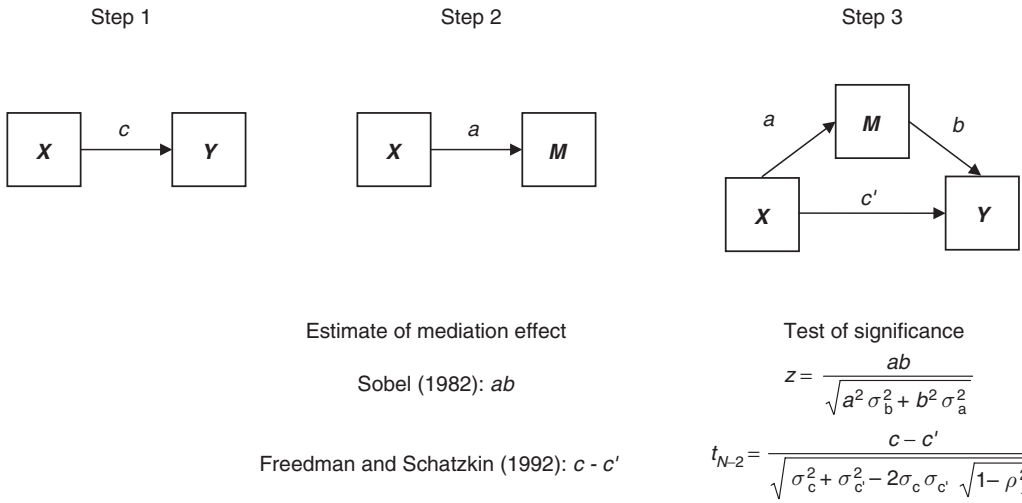
HLM-Based Tests of Multilevel Mediation

Investigating mediators and understanding the mediation processes underlying observed relationships are ''what moves organizational research beyond dust-bowl empiricism and toward a true science'' (Mathieu et al., 2008, p. 203). Mediators are ''variables through which the influence of an antecedent variable is transferred to a criterion'' (Mathieu & Taylor, 2007, p. 142). In single-level settings, Baron and Kenny (1986) proposed several conditions to be met for a variable to function as a mediator:

> ''(a) variations in the levels of the independent variable significantly account for *variations* in the presumed mediator (i.e., Path *a*), (b) variations in the mediator significantly account for *variations* in the dependent variable (i.e., Path *b*), and (c) when Path *a* and *b* are controlled, a previously significant relation between the independent and dependent variables is no longer significant, with the strongest demonstration of mediation occurring when Path *c* is zero'' (p. 1176, emphasis added; See Figure 1 for Paths *a, b*, and *c*).

As with other methods for testing mediation, Baron and Kenny's (1986) procedures have been reformulated in multilevel settings (e.g., Krull & MacKinnon, 2001; Mathieu & Taylor,

Figure 1
Baron and Kenny's (1986) Steps for Single-Level Mediation Analysis and Associated Tests



Note: σ refers to the standard error of the corresponding parameter and ρ_{XM} refers to the correlation coefficient between the independent variable X and the mediator variable M.

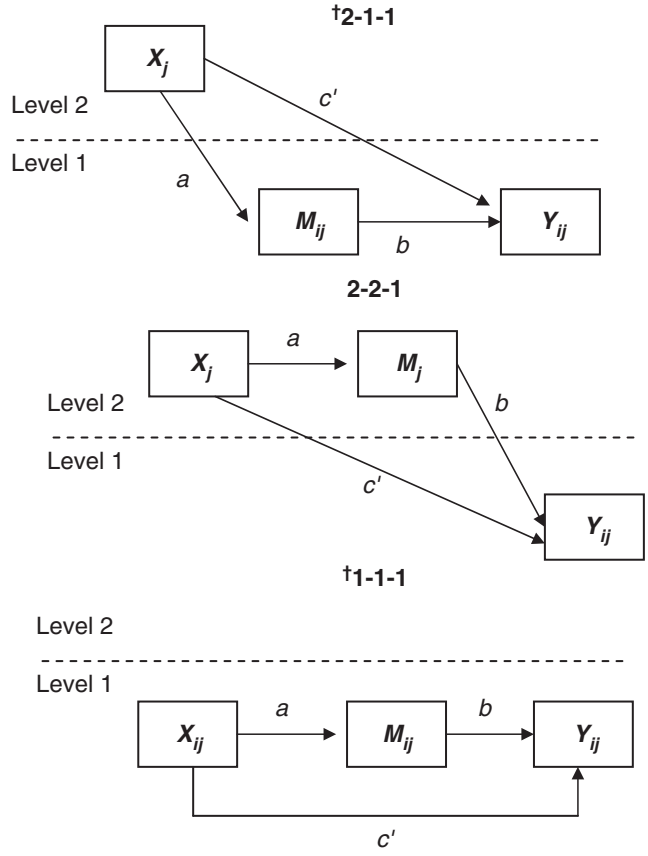
2007). Figure 2 provides an illustration of three multilevel mediation models typically found in research. Labels of the models are based on the level of measurement of the antecedent (X), the mediator (M), and the outcome variable (Y). In particular, if all three variables are measured at Level-1, the model is labeled as 1-1-1; if the antecedent is measured at Level-2, while the mediator and outcome are at Level-1, it is labeled as 2-1-1; if the antecedent and mediator are both measured at Level-2 and the outcome at Level-1, it is labeled as 2-2-1. Similar labels have been used by Bauer, Preacher, and Gil (2006) and Krull and MacKinnon (2001).

In the 2-1-1 model, a Level-2 antecedent influences a Level-1 mediator which then affects a Level-1 outcome. An example of this mediation mechanism can be found in Seibert, Silver, and Randolph's (2004) study examining the relationship between empowerment climate (a Level-2 antecedent) and job satisfaction (a Level-1 outcome), as mediated by individuals' psychological empowerment.

In the 2-2-1 model, a Level-2 antecedent influences a Level-2 mediator, which in turn affects a Level-1 outcome. An example of this mediation mechanism can be found in Chen, Kirkman, Kanfer, Allen, and Rosen's (2007) study investigating the influence of leadership climate (a Level-2 antecedent) on individual empowerment (a Level-1 outcome) through team empowerment (a Level-2 mediator).

In the 1-1-1 model, the antecedent, the mediator, and the outcome are all measured at a lower level of analysis, but the Level-1 units (e.g., individuals or firms) are nested in Level-2 units (e.g., teams or industries). An example of this mediation mechanism would be the effect of individual procedural fairness perceptions (a Level-1 antecedent) on individual

Figure 2
Three Types of HLM-Based Multilevel Mediation Models



Note: †indicates that grand-mean centered HLM models can produce confounded estimates of the mediation effects. HLM = hierarchical linear modeling.

citizenship behaviors (a Level-1 outcome) through organizational identification (a Level-1 mediator). Such a causal chain is a multilevel phenomenon when the sample consists of employees from dozens of firms (i.e., Level-2 units), and the firm–level relationships may be different from individual-level relationships (Klein & Kozlowski, 2000; Ostroff, 1993).

Although there are many ways to quantify mediation effects, we limit discussion to two in this study: The *product-of-coefficients* method (operationalized as *ab*) and the *difference-in-coefficients* (of *X* affecting *Y*) method (operationalized as $c - c'$). The paths *a, b*, and c' are shown in Figure 1, and *c* refers to the path from *X* to *Y* when *M* is absent. The $\hat{c} - \hat{c}'$ and $\hat{a}\hat{b}$ sample estimates are algebraically equivalent in single-level models (MacKinnon, Warsi, & Dwyer, 1995) and $\hat{c} - \hat{c}' = \hat{a}\hat{b}$ is considered ''a fundamental equality at the heart of the vast majority of mediational analysis'' (Muller, Yzerbyt, & Judd, 2008, p. 225). However, in multilevel models, $\hat{c} - \hat{c}'$ and $\hat{a}\hat{b}$ can produce different values. These two quantities are

interpreted differently when multiple mediators are present: $\hat{c} - \hat{c}'$ can estimate the total mediation effect, whereas $\hat{a}\hat{b}$ estimates a unique mediation effect for a single mediator (Krull & MacKinnon, 1999).

We discuss these two methods because of their popularity in mediation testing. Both methods were discussed by Baron and Kenny (1986), and the product-of-coefficients method is usually used when quantifying mediation effects. Researchers have suggested that bootstrapping and the empirical-*M* test (see MacKinnon, Lockwood, & Williams, 2004) are preferable for testing multilevel mediation effects, especially when the variables are not normally distributed (e.g., Pituch & Stapleton, 2008). However, the purpose of our study is not to compare the statistical performance of various methods for detecting mediation effects. Instead, we endeavor to demonstrate (and offer corrections for) potential confounding in commonly used estimation procedures and the inaccuracy of associated test statistics.

Available Procedures for Testing Multilevel Mediation

Table 1 provides a summary of the available procedures for testing multilevel mediation for the three types of models. It is notable that previous research has given no preference to a specific centering approach. For example, Hofmann and Gavin (1998) suggested that either grand-mean centering or group-mean centering with the means reintroduced into the Level-2 intercept model would provide an appropriate test of mediation effects. However, as the following sections of this article show, grand-mean centering or no centering may produce confounded point estimates of the mediation effect.

The following discussion uses the 2-1-1 model as an example of a multilevel mediation model. To facilitate our discussion and interpretations, the following example is helpful: Does job satisfaction (a Level-1 mediator) mediate the relationship between flexible work schedule (a Level-2 antecedent) and employee performance (a Level-1 outcome)? Assume data were collected on job satisfaction and performance from employees working in dozens of firms, where each firm has provided information on the extent to which it has implemented flexible work schedules for employees. This firm–level antecedent is treated as a continuous variable. In the following discussion, we treat firms and groups as interchangeable when we refer to the Level-2 units.

To answer the above research question, the first step in testing the mediation effect is often to establish a relationship between flexible work schedule (the Level-2 antecedent X_j) and employee performance (the Level-1 outcome Y_{ij}). Equation (1) corresponds to the Level-1 equation for employee performance and Equation (2) corresponds to the Level-2 equation for the intercept of the Level-1 equation:

$$\text{Level 1: } Y_{ij} = \beta_{0j}^{(1)} + r_{ij}^{(1)} \tag{1}$$

$$\text{Level 2: } \beta_{0j}^{(1)} = \gamma_{00}^{(1)} + \gamma_{01}^{(1)} X_j + u_{0j}^{(1)} \tag{2}$$

where subscripts *i* and *j* refer to individuals and Level-2 units (e.g., firms), respectively; β_{0j} is the intercept for firm *j*; r_{ij} and u_{0j} are the Level-1 and Level-2 residuals, respectively. The superscript 1 denotes coefficients, parameters, and random variables for the first set of equations.

Table 1
Comparison of HLM-Based Multilevel Mediation Models (Two-Level Models Only)

Baron and Kenny (1986) Step 1	Step 2	Step 3
2-1-1 Model		
Available procedure (Grand-mean centering or no centering)		
Mediation effect: $\gamma_{01}^{(2)} * \gamma_{10}^{(3)}$ or $\gamma_{01}^{(1)} - \gamma_{01}^{(3)}$		
L 1: $Y_{ij} = \beta_{0j}^{(1)} + r_{ij}^{(1)}$ (1)	L 1: $M_{ij} = \beta_{0j}^{(2)} + r_{ij}^{(2)}$ (3)	L 1: $Y_{ij} = \beta_{0j}^{(3)} + \beta_{1j}^{(3)} M_{ij} + r_{ij}^{(3)}$ (5)
L 2: $\beta_{0j}^{(1)} = \gamma_{00}^{(1)} + \gamma_{01}^{(1)} X_j + u_{0j}^{(1)}$ (2)	L 2: $\beta_{0j}^{(2)} = \gamma_{00}^{(2)} + \gamma_{01}^{(2)} X_j + u_{0j}^{(2)}$ (4)	L 2: $\beta_{0j}^{(3)} = \gamma_{00}^{(3)} + \gamma_{01}^{(3)} X_j + u_{0j}^{(3)}$ (6)
		$\beta_{1j}^{(3)} = \gamma_{10}^{(3)}$ (7)
Proposed procedure (group-mean centering and adding the group mean at level 2, or CWC(M))		
Mediation effect: $\gamma_{01}^{(2)} * \gamma_{02}^{(4)}$ or $\gamma_{01}^{(1)} - \gamma_{01}^{(4)}$		
L 1: Equation (1)	L 1: Equation (3)	L 1: $Y_{ij} = \beta_{0j}^{(4)} + \beta_{1j}^{(4)} (M_{ij} - M_{\bullet j}) + r_{ij}^{(4)}$ (8)
L 2: Equation (2)	L 2: Equation (4)	L 2: $\beta_{0j}^{(4)} = \gamma_{00}^{(4)} + \gamma_{01}^{(4)} X_j + \gamma_{02}^{(4)} M_{\bullet j} + u_{0j}^{(4)}$ (9)
		$\beta_{1j}^{(4)} = \gamma_{10}^{(4)}$ (10)
2-2-1 Model		
Grand-mean-centering and CWC(M) are equivalent		
Mediation effect: $\gamma_{01}^{(2)} * \gamma_{02}^{(3)}$ or $\gamma_{01}^{(1)} - \gamma_{01}^{(3)}$ (Unconfounded estimate of mediation effects)		
L 1: $Y_{ij} = \beta_{0j}^{(1)} + r_{ij}^{(1)}$ (11)	L 2: $M_j = \gamma_{00}^{(2)} + \gamma_{01}^{(2)} X_j + u_{0j}^{(2)}$ (13)	L 1: $Y_{ij} = \beta_{0j}^{(3)} + r_{ij}^{(3)}$ (14)
L 2: $\beta_{0j}^{(1)} = \gamma_{00}^{(1)} + \gamma_{01}^{(1)} X_j + u_{0j}^{(1)}$ (12)		L 2: $\beta_{0j}^{(3)} = \gamma_{00}^{(3)} + \gamma_{01}^{(3)} X_j + \gamma_{02}^{(3)} M_j + u_{0j}^{(3)}$ (15)

Table 1 (continued)

Baron and Kenny (1986) Step 1	Step 2	Step 3
1-1-1 Model		
Available procedure (grand-mean centering or no centering)		
Mediation effect: $\gamma_{10}^{(2)} * \gamma_{20}^{(3)}$ or $\gamma_{10}^{(1)} - \gamma_{10}^{(3)}$ (Level-1 effect confounded with Level-2 effect)		
L 1: $Y_{ij} = \beta_{0j}^{(1)} + \beta_{1j}^{(1)} X_{ij} + r_{ij}^{(1)}$ (16)	L 1: $M_{ij} = \beta_{0j}^{(2)} + \beta_{1j}^{(2)} X_{ij} + r_{ij}^{(2)}$ (19)	L 1: $Y_{ij} = \beta_{0j}^{(3)} + \beta_{1j}^{(3)} X_{ij} + \beta_{2j}^{(3)} M_{ij} + r_{ij}^{(3)}$ (22)
L 2: $\beta_{0j}^{(1)} = \gamma_{00}^{(1)} + u_{0j}^{(1)}$ (17)	L 2: $\beta_{0j}^{(2)} = \gamma_{00}^{(2)} + u_{0j}^{(2)}$ (20)	L 2: $\beta_{0j}^{(3)} = \gamma_{00}^{(3)} + u_{0j}^{(3)}$ (23)
$\beta_{1j}^{(1)} = \gamma_{10}^{(1)}$ (18)	$\beta_{1j}^{(2)} = \gamma_{10}^{(2)}$ (21)	$\beta_{1j}^{(3)} = \gamma_{10}^{(3)}$ (24)
		$\beta_{2j}^{(3)} = \gamma_{20}^{(3)}$ (25)
Proposed procedure CWC(M)		
Mediation effects: Level 1 effect $\gamma_{10}^{(5)} * \gamma_{20}^{(6)}$ or $\gamma_{10}^{(4)} - \gamma_{10}^{(6)}$; level 2 effect $\gamma_{01}^{(5)} * \gamma_{02}^{(6)}$ or $\gamma_{01}^{(4)} - \gamma_{01}^{(6)}$		
L 1: $Y_{ij} = \beta_{0j}^{(4)} + \beta_{1j}^{(4)} (X_{ij} - X_{\bullet j}) + r_{ij}^{(4)}$ (26)	L 1: $M_{ij} = \beta_{0j}^{(5)} + \beta_{1j}^{(5)} (X_{ij} - X_{\bullet j}) + r_{ij}^{(5)}$ (29)	L 1: $Y_{ij} = \beta_{0j}^{(6)} + \beta_{1j}^{(6)} (X_{ij} - X_{\bullet j}) + \beta_{2j}^{(6)} (M_{ij} - M_{\bullet j}) + r_{ij}^{(6)}$ (32)
L 2: $\beta_{0j}^{(4)} = \gamma_{00}^{(4)} + \gamma_{01}^{(4)} X_{\bullet j} + u_{0j}^{(4)}$ (27)	L 2: $\beta_{0j}^{(5)} = \gamma_{00}^{(5)} + \gamma_{01}^{(5)} X_{\bullet j} + u_{0j}^{(5)}$ (30)	L 2: $\beta_{0j}^{(6)} = \gamma_{00}^{(6)} + \gamma_{01}^{(6)} X_{\bullet j} + \gamma_{02}^{(6)} M_{\bullet j} + u_{0j}^{(6)}$ (33)
$\beta_{1j}^{(4)} = \gamma_{10}^{(4)}$ (28)	$\beta_{1j}^{(5)} = \gamma_{10}^{(5)}$ (31)	$\beta_{1j}^{(6)} = \gamma_{10}^{(6)}$ (34)
		$\beta_{2j}^{(6)} = \gamma_{20}^{(6)}$ (35)

Note: 1-1-1 Model refers to the model where *X, M*, and *Y* were all measured at level 1; 2-1-1 model refers to the model where *X* is level 2 and *M* and *Y* are both measured at level 1; 2-2-1 Model refers to the model where *X* and *M* are both at level 2 and *Y* is at level 1. CWC(M) = centered within context with reintroduction of the subtracted means; HLM = hierarchical linear modeling.

The second step is often to show a relationship between a firm's flexible work schedule X_j and the individual-level mediator job satisfaction (M_{ij}). The equations in this step (Equations [3] and [4]) are shown in Table 1. If grand-mean centering is used for job satisfaction and work performance, the third step is to show that after adding job satisfaction to the model at Level-1, the effect of flexible work schedule on work performance is reduced in magnitude, whereas job satisfaction is still a statistically significant predictor of the outcome.

$$\text{Level 1: } Y_{ij} = \beta_{0j}^{(3)} + \beta_{1j}^{(3)} M_{ij} + r_{ij}^{(3)} \tag{5}$$

$$\text{Level 2: } \beta_{0j}^{(3)} = \gamma_{00}^{(3)} + \gamma_{01}^{(3)} X_j + u_{0j}^{(3)} \tag{6}$$

$$\beta_{1j}^{(3)} = \gamma_{10}^{(3)} \tag{7}$$

It is notable that in Equations (5–7), the mediator job satisfaction could have a random slope. Although adding a random effect to the slope equation is possible, this may add unnecessary complications to the models, resulting in increased rates of nonconvergence. As we foresee no difficulty in generalizing our findings to models containing random slopes, we do not complicate the current discussion with the addition of random slopes.

According to Freedman and Schatzkin (1992), a significant decrease in the coefficients of flexible work schedule (X_j) in Equation (6), as compared to Equation (2), would indicate a mediation effect of job satisfaction (M_{ij}) in the relationship between flexible work schedule and performance (Y_{ij}). Under grand-mean centering, this mediation effect is represented by $\gamma_{01}^{(1)} - \gamma_{01}^{(3)}$. Freedman and Schatzkin (1992) have suggested a t statistic to test the significance of the reduction in the coefficients (see Figure 1 for the formula associated with this test). Using the product-of-coefficients method, the product $\gamma_{01}^{(2)} * \gamma_{10}^{(3)}$ represents the mediation effect, and a Sobel z statistic (Sobel, 1982) can be used to test the significance of this effect.

Potential Confounding in Grand-Mean-Centered 2-1-1 Models

The above grand-mean-centered HLM procedures do not consider the unique data structure in multilevel models and thus could raise the possibility of confounded mediation-effect estimations. In particular, the relationship between two Level-1 variables (job satisfaction and work performance, in the above example) can be decomposed into between-group and within-group components (Raudenbush & Bryk, 2002)—we use the term ''group'' broadly to represent the Level-2 units in which individuals are nested. Davis, Spaeth, and Huson (1961) discussed the difference between within-group and between-group regressions in multilevel settings and noted that the coefficients of these two types of regressions can be very different, and even opposite in sign (any difference in these coefficients is known as a ''contextual effect''; Raudenbush & Bryk, 2002). In particular, the aggregate amount of job satisfaction at the group level may strongly predict group-level performance whereas the within-group relationship between job satisfaction and performance may be very weak.

Differentiating these two components allows answering very different questions than when they are not differentiated. Another example is the relationship between student socioeconomic status (SES) and scholarly performance. The effect of the school-average SES on average performance can be very different from the effect of within-school SES (measured as individual deviation from the school average). Here, the between-group and within-group relationships could even have opposite signs (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). Similar arguments have been provided by proponents of the Within-And-Between-Entity analysis (WABA) technique (see Dansereau et al., 1984). In the classical WABA equation, the raw score correlation between any two Level-1 variables is decomposed into its between-group and within-group components, and the substantive meaning of the between-group relationship is not necessarily the same as that of the within-group relationship.

In the context of multilevel mediation tests based on 2-1-1 models, both within- and between-group effects may be contained in a single mediation effect estimate. Continuing with our example of flexible work schedule and work performance, we now show how potential confounding could arise. Instead of using Equations (5–7) with grand-mean centering, now we use the following equations, in which the group-mean centered job satisfaction is used and its group (or firm) mean ($M_{\bullet j}$) is included at Level-2.

$$\text{Level 1: } Y_{ij} = \beta_{0j}^{(4)} + \beta_{1j}^{(4)}(M_{ij} - M_{\bullet j}) + r_{ij}^{(4)} \tag{8}$$

$$\text{Level 2: } \beta_{0j}^{(4)} = \gamma_{00}^{(4)} + \gamma_{01}^{(4)}X_j + \gamma_{02}^{(4)}M_{\bullet j} + u_{0j}^{(4)} \tag{9}$$

$$\beta_{1j}^{(4)} = \gamma_{10}^{(4)} \tag{10}$$

where $M_{\bullet j}$ is the average value of job satisfaction for group j and $\gamma_{02}^{(4)}$ and $\gamma_{10}^{(4)}$ are the between-group and within-group coefficients of job satisfaction, respectively. Kreft and de Leeuw (1998, p. 110) termed the group-mean centered analysis ''centered within context'' or CWC, and they specifically explore CWC(M) (centered within context with reintroduction of the subtracted means at Level-2; see Equations [8–10]). Comparing grand-mean centering (Equations [5–7]) with CWC(M), the difference in these models reduces to the fact that the within-group coefficient of job satisfaction (i.e., $\gamma_{10}^{(4)}$) is held equal to the between-group coefficient of job satisfaction (i.e., $\gamma_{02}^{(4)}$) in grand-mean centering (Snijders & Bosker, 1999). To show how this occurs, we first substitute $\beta_{0j}^{(3)}$ and $\beta_{1j}^{(3)}$ in Equation (5) with their expressions in Equations (6) and (7), and then rewrite $\gamma_{10}^{(3)} * M_{ij}$ into $\gamma_{10}^{(3)} * [(M_{ij} - M_{\bullet j}) + M_{\bullet j}] = \gamma_{10}^{(3)} * (M_{ij} - M_{\bullet j}) + \gamma_{10}^{(3)} * M_{\bullet j}$. Now, in comparison with Equations (8–10), grand-mean centering places a possibly unwarranted constraint on the model, such that $\gamma_{10}^{(4)}$ and $\gamma_{02}^{(4)}$ are made equal and both are equal to the single coefficient $\gamma_{10}^{(3)}$.

Owing to this constraint, Equations (5–7) may provide confounded and incorrect estimates of the mediation effect if researchers are solely interested in the Level-2 relationships when they examine the 2-1-1 model. Recall that we are interested in the mediating role of job satisfaction in the relationship between flexible work schedule and employee performance. Because flexible work schedule varies only between Level-2

units, it cannot be associated with differences across people within firms. Any effect of flexible work schedule on performance, whether mediated or not, can exist only between firms. Thus, in this situation, researchers should focus on the between-group mediation effect (rather than a combination of between- and within-group mediation effects). Because the within-group relationship under CWC(M) is independent of the between-group relationship (e.g., Kreft & de Leeuw, 1998), a mediation estimate combining the two will serve to make ambiguous the 2-1-1 mediation effect.

Given the above argument, it may not be appropriate to use the coefficient $\gamma_{10}^{(3)}$ (in Equation [7]) to calculate a mediation effect because it represents a ''forced average'' of the within-group effect and the between-group effect of job satisfaction on work performance. Under CWC(M), the reduction in the coefficients of flexible work schedule is $\gamma_{01}^{(1)} - \gamma_{01}^{(4)}$ and the product of coefficients is $\gamma_{01}^{(2)} * \gamma_{02}^{(4)}$. In contrast, under grand-mean centering, the estimates are $\gamma_{01}^{(1)} - \gamma_{01}^{(3)}$ and $\gamma_{01}^{(2)} * \gamma_{10}^{(3)}$, respectively. Because the possibly unwarranted constraint is added under grand-mean centering, $\gamma_{01}^{(4)}$ may not be equal to $\gamma_{01}^{(3)}$, and $\gamma_{02}^{(4)}$ may not be equal to $\gamma_{10}^{(3)}$.

Admittedly, the within-group relationship between job satisfaction and work performance may also be of interest to researchers. Klein and Kozlowski (2000) have referred to these within-group-relationship models as frog-pond models (p. 219). A frog-pond model examines the relationship between a group member's relative standing within a group on job satisfaction and his or her performance. Unfortunately, because the Level-2 antecedent cannot account for the relative standing of group members in terms of job satisfaction (i.e., flexible work practices that are invariant within firms cannot be conceptualized as causing differences across individuals within a given firm in job satisfaction or performance), the within-group coefficient linking job satisfaction and performance is distracting in a 2-1-1 model. Such an effect is likely to be, and perhaps should be, irrelevant in any 2-1-1 model where the focus is on mediation of the effect of a Level-2 antecedent variable.

Using the study by Seibert et al. (2004) as an example, researchers can obtain greater insight into mediational mechanisms if the between- and within-group relationships are separately estimated in 2-1-1 models. In particular, their study examines the mediating role of individuals' psychological empowerment (a Level-1 mediator) in the relationship between empowerment climate (a Level-2 antecedent) and job satisfaction (a Level-1 outcome). Three different mechanisms could result in the same conclusion of a mediation effect when between- and within-group effects are confounded. First, mediation could be mainly at the between-group level, such that there is a strong between-group relationship but a weak within-group relationship between psychological empowerment and job satisfaction. In this case, aggregated psychological empowerment mediates the relationship between empowerment climate and aggregated job satisfaction. Second, there may be a strong within-group empowerment-satisfaction relationship but a weak between-group relationship. Third, the relationships at two levels may both be moderate in strength. The above three mechanisms may answer different research questions, and they can be differentiated using CWC(M).

Two factors may influence the severity of the confounding when grand-mean or no centering is used. The first is the magnitude of the within-group effect of the mediator

on the outcome (i.e., $\gamma_{10}^{(4)}$) relative to the between-group effect (i.e., $\gamma_{02}^{(4)}$). When the magnitude of the within-group effect departs from the magnitude of the between-group effect, the estimates of the mediation effect could be underestimated or overestimated when using grand-mean centering, because the total effect of job satisfaction on work performance (i.e., $\gamma_{10}^{(3)}$) may not accurately represent the between-group effect. Instead, the total mediation effect derived in Equations (1–7) occurs in part as a function of the within-group effect.

The second factor influencing the severity of the confounding in testing multilevel mediation is the group size (i.e., the number of individuals within the groups), relative to the total sample size. With few groups and larger group sizes, the impact of the within-group portion of job satisfaction on work performance ($\gamma_{10}^{(3)}$) is likely to increase; alternatively, with many groups and fewer people in each group, $\gamma_{10}^{(3)}$ will be less influenced by the within-group part of the relationship. In other words, any factors that make the between-group component more influential could increase the accuracy of the point estimate of a mediation effect if researchers are interested in the mediation effect that exists between groups.

To examine the influence of the magnitude of between- versus within-group effects and group size on the extent of confounding in 2-1-1 mediation testing, a Monte Carlo simulation study was conducted. The details of this study are provided in the following section. It is notable that similar confounding may occur in the more complicated 1-1-1 models if researchers are only interested in the Level-2 or the Level-1 mediation effects. In these situations, simply combining the two mediation effects at the two levels could lead to confounded estimates of both effects. We discuss the implications of the simulation results for the 1-1-1 model in our discussion section. In contrast to 1-1-1 and 2-1-1 models, 2-2-1 models do not suffer from this type of confounding in estimating mediation effects, because the relationships between the antecedent and the outcome and between the mediator and the outcome are all at Level 2. Consequently, there are no Level-1 relationships that could interfere with the estimation of Level-2 mediation effects.

Simulation Study of a 2-1-1 Model

A Monte Carlo simulation was conducted to examine the extent of confounding using the grand-mean-centered HLM procedure to estimate the 2-1-1 mediation effect.

Model Specification and Data Generation

The Stata (version 9.0) programming architecture was used for all simulations and analyses. Simulated data were generated based on Equations (8–10) and analyzed using both the traditional procedure and CWC(M). This study was a 4 (between-group coefficient $\gamma_{02}^{(4)} = 0$, .14, .39, .59) by 7 (within-group coefficient $\gamma_{10}^{(4)} = -.59, -.39, -.14, 0, .14, .39, .59$) by 4 (group size $n = 5, 8, 12, 20$ with total sample size fixed at 600) factorial design. Each cell of the design contained 500 samples (i.e., replications). The Stata syntax for data generation and analyses is available at http://www.quantpsy.org.

The group size ($n = 5, 8, 12, 20$) was chosen to be comparable to common group sizes in the social sciences (e.g., Krull & MacKinnon, 1999). The parameter values (0, .14, .39, .59)

were chosen to correspond to zero, small, medium, and large effect sizes (see Cohen, 1988, pp. 412-414) and to correspond as closely as possible with previous simulation research (e.g., Cheung & Lau, 2008; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). Decisions with regard to the distributions of the random effects and effect sizes of the fixed parameters are determined as follows. In Equations (8–10), the Level-2 grand intercept ($\gamma_{00}^{(4)}$) was specified as 0 and $\gamma_{10}^{(4)}$ was fixed at .30; setting the grand-mean equal to 0 simplifies the model; a coefficient of .30 has been used to represent a medium effect in previous multilevel simulation studies (e.g., Krull & MacKinnon, 1999). The independent variable X_j and the between-group mediator $M_{\bullet j}$ follow a standard bivariate normal distribution at Level-2, with a bivariate correlation of .40. The variance of the Level-1 error term is .25 and the variance of the Level-2 error term is .05, so that the conditional intraclass correlation (ICC) = .17. This medium ICC value was chosen to facilitate model convergence—previous research has shown convergence problems when the residual ICC is small (e.g., Busing, 1993).

Analysis

After the data sets were generated, they were first analyzed using Equations (1–7) and then analyzed again using Equations (1–4) and (8–10). Results of the grand-mean-centered multilevel models were compared with those of the CWC(M) models. In particular, the point estimates of the mediation effect (in terms of the product of coefficients $\hat{a}\hat{b}$ and difference in coefficients $\hat{c} - \hat{c}'$) and their corresponding test statistics were compared.

Extent of confounding. The extent of confounding in the point estimates for the mediation effect was computed by first subtracting the average sample value based on CWC(M) from the average sample value based on grand-mean centering across the 500 replications for each condition, then dividing by the average sample value of CWC(M), and then multiplying by 100%. The extent of confounding is not reported when $\gamma_{02}^{(4)}$ is expected to be zero.

$$\text{Extent of confounding for } \hat{a}\hat{b} = \frac{E[\gamma_{01}^{(2)} * \gamma_{10}^{(3)}] - E[\gamma_{01}^{(2)} * \gamma_{02}^{(4)}]}{E[\gamma_{01}^{(2)} * \gamma_{02}^{(4)}]} \times 100\% \tag{36}$$

Inaccuracy of the Sobel statistic was calculated by subtracting the average CWC(M) Sobel statistic from the average grand-mean-centered value, and dividing this value by the average CWC(M) Sobel statistic across 500 replications.

$$\text{Inaccuracy of Sobel statistic} = \frac{E(\text{Sobel}_{\text{Grand-Mean}}) - E(\text{Sobel}_{\text{CWC(M)}})}{E(\text{Sobel}_{\text{CWC(M)}})} \times 100\% \tag{37}$$

$$\text{Extent of confounding for } \hat{c} - \hat{c}' = \frac{E[\gamma_{01}^{(1)} - \gamma_{01}^{(3)}] - E[\gamma_{01}^{(1)} - \gamma_{01}^{(4)}]}{E[\gamma_{01}^{(1)} - \gamma_{01}^{(4)}]} \times 100\% \tag{38}$$

$$\text{Inaccuracy of Freedman \& Schatzkin (1992) statistic} = \frac{E(\text{FS}_{\text{Grand-Mean}}) - E(\text{FS}_{\text{CWC(M)}})}{E(\text{FS}_{\text{CWC(M)}})} \times 100\% \quad (39)$$

Calculation of empirical power and Type-I error rate. Empirical power is calculated in the $\gamma_{02}^{(4)} = .14$ situation, because we are interested in the empirical power to detect a small effect size and .14 represents a small effect size (MacKinnon et al., 2002). Empirical power of the statistical tests is the proportion of replications for which the null hypothesis (i.e., that the parameter of interest, $c - c'$ or ab, equals zero) is rejected at the $\alpha = .05$ level. When $\gamma_{02}^{(4)} = 0$, Type-I error rates are reported.

Discrepancy between $\hat{c} - \hat{c}'$ and $\hat{a}\hat{b}$. Previous simulations found that under grand-mean centering, mean discrepancies between $\hat{c} - \hat{c}'$ and $\hat{a}\hat{b}$ are zero (Krull & MacKinnon, 1999). We examine the discrepancy (in percentage) between these two statistics under CWC(M).

Results

Extent of Confounding in Mediation Effect Estimates

Extent of confounding for the mediation effects $\hat{a}\hat{b}$ and $\hat{c} - \hat{c}'$ are provided in Tables 2 and 3, respectively. With regard to the product-of-coefficients method, the point estimates under grand-mean centering show a large degree of inaccuracy in most conditions. The confounded effects show a pattern of positive versus negative signs: When the within-group coefficient is smaller or larger than the between-group coefficient, the point estimate of the mediation effect and the associated statistic show a negative or positive difference, respectively. As the within-group coefficient becomes more similar to the between-group coefficient, grand-mean centering provides more accurate estimates.

This pattern is expected, as grand-mean centering produces an ''average'' of the between- and within-group coefficients. Thus, when the within-group effect is greater than the between-group effect, the confounded mediation estimate is greater than the true effect (i.e., the between-group effect). When the within-group effect is less than the between-group effect, the confounded estimate is smaller than the true effect. Comparing Tables 2 and 3, similar patterns emerge for $\hat{a}\hat{b}$ and $\hat{c} - \hat{c}'$. Both methods are subject to severe confounding when the between- and within-group coefficients differ in magnitude. Note that in these two tables, the extent of confounding is not reported when the true between-group coefficient is .00, because the denominator would be zero.

Tables 2 and 3 show that, all else being equal, when group size increases (and thus the number of groups decreases) the within-group coefficient gains weight, increasing the observed confounding. Tables 2 and 3 also show that the Sobel statistic and Freedman and Schatzkin (1992) statistic are somewhat inefficient. When the between- and within-group coefficients take the same value, these statistics are relatively accurate (i.e., inaccuracy is

Table 2
Inaccuracy (in Percentage) of Mediation Effect Estimates $\hat{ab}$ and Sobel Statistic Estimates

		−.59		−.39		−.14		0		.14		.39		.59	
Within-Group Coefficient $\gamma_{10}^{(4)}$		$\hat{ab}$	S	$\hat{ab}$	S	$\hat{ab}$	S	$\hat{ab}$	S	$\hat{ab}$	S	$\hat{ab}$	S	$\hat{ab}$	S
Size	# Groups	True between-group coefficient $\gamma_{02}^{(4)}$ is .14													
5	120	−468.0	−242.9	−315.1	−238.5	−148.0	−188.4	−70.9	−36.7	−.60	25.4	128.2	41.3	260.2	44.9
8	75	−496.9	−245.4	−342.2	−240.2	−161.7	−207.8	−77.0	−42.0	.89	28.7	142.0	38.8	279.4	41.8
12	50	−490.7	−244.7	−352.6	−242.3	−173.6	−220.3	−82.8	−51.7	2.0	33.9	149.7	43.0	291.5	41.2
20	30	−518.3	−251.4	−362.4	−249.7	−182.9	−233.7	−87.7	−60.5	−1.8	39.3	162.4	51.3	318.4	49.4
		True between-group coefficient $\gamma_{02}^{(4)}$ is .39													
5	120	−237.0	−205.2	−182.4	−202.2	−114.3	−153.5	−78.0	−27.8	−46.5	−1.29	.27	5.01	37.0	5.9
8	75	−243.6	−205.9	−190.0	−203.9	−123.2	−178.4	−85.9	−40.0	−51.6	−1.05	−.25	5.3	40.6	6.3
12	50	−244.6	−206.3	−193.5	−205.2	−127.2	−189.1	−89.8	−48.1	−55.0	−.19	−.63	5.7	42.7	6.4
20	30	−248.1	−207.5	−196.8	−207.3	−131.0	−196.8	−94.6	−66.2	−58.1	2.0	−1.3	7.0	46.6	7.4
		True between-group coefficient $\gamma_{02}^{(4)}$ is .59													
5	120	−191.1	−201.3	−156.0	−198.5	−111.4	−156.9	−86.7	−36.3	−61.6	−4.9	−24.2	1.3	.19	2.4
8	75	−194.8	−202.1	−160.8	−200.4	−116.3	−177.4	−91.7	−47.7	−67.1	−5.1	−26.7	1.6	−.27	2.3
12	50	−197.5	−202.6	−162.7	−201.2	−119.2	−186.2	−94.8	−59.7	−70.6	−4.5	−28.2	2.1	−.16	2.7
20	30	−198.9	−203.5	−164.4	−202.6	−120.8	−192.5	−97.1	−73.4	−72.5	−2.9	−30.9	3.0	−.01	3.9

Note. Total $N = 600$. $\hat{ab}$ refers to $\gamma_{01}^{(2)} * \gamma_{10}^{(3)}$ and S refers to Sobel statistic. Extent of confounding in the mediation effect was calculated based on Equation 36; Inaccuracy of the Sobel statistic was calculated based on Equation 37.

Table 3
Inaccuracy (in Percentage) of Mediation Effect Estimates $\hat{c} - \hat{c}'$ and Freedman and Schatzkin (1992) Statistic Estimates

Within-Group Coefficient $\gamma_{10}^{(4)}$		−.59		−.39		−.14		0		.14		.39		.59	
		$\hat{c} - \hat{c}'$	FS	$\hat{c} - \hat{c}'$	FS	$\hat{c} - \hat{c}'$	FS	$\hat{c} - \hat{c}'$	FS	$\hat{c} - \hat{c}'$	FS	$\hat{c} - \hat{c}'$	FS	$\hat{c} - \hat{c}'$	FS
Size	# groups	True between-group coefficient $\gamma_{02}^{(4)}$ is .14													
5	120	−456.2	−264.0	−316.1	−223.8	−147.6	−142.6	−70.9	−70.3	−.60	.35	127.8	139.1	256.0	280.6
8	75	−510.3	−240.6	−340.8	−205.7	−161.6	−152.0	−76.9	−76.1	.88	2.1	148.3	153.4	279.0	214.2
12	50	−492.9	−212.5	−360.5	−196.1	−173.2	−158.2	−82.6	−81.7	2.2	4.4	152.7	149.1	291.8	158.9
20	30	−513.3	−193.2	−362.2	−192.2	−182.2	−163.3	−87.8	−87.1	−1.5	3.0	158.2	140.1	319.5	128.5
		True between-group coefficient $\gamma_{02}^{(4)}$ is .39													
5	120	−236.1	−176.4	−182.7	−158.7	−114.3	−116.7	−78.0	−72.0	−46.6	−42.6	.23	−5.8	36.9	41.0
8	75	−243.4	−168.2	−188.6	−155.3	−123.3	−126.2	−85.9	−80.5	−51.4	−45.3	−.25	−6.8	40.3	49.3
12	50	−245.9	−164.5	−194.7	−155.3	−127.3	−130.7	−89.8	−85.2	−55.0	−48.0	−.56	−6.8	42.4	55.7
20	30	−245.1	−158.7	−197.1	−154.3	−131.0	−132.6	−94.6	−92.0	−57.9	−49.8	−1.3	−7.5	46.5	66.3
		True between-group coefficient $\gamma_{02}^{(4)}$ is .59													
5	120	−190.7	−168.0	−155.4	−153.3	−111.4	−117.6	−86.7	−77.3	−61.5	−47.5	−24.5	−23.8	.18	−4.7
8	75	−194.4	−164.4	−160.1	−153.9	−116.2	−125.0	−91.7	−84.6	−67.1	−51.6	−26.8	−24.9	−.29	−5.14
12	50	−196.5	−162.4	−162.2	−154.8	−119.2	−128.7	−94.8	−89.9	−70.7	−54.4	−28.1	−26.1	−.21	−5.6
20	30	−199.3	−160.1	−164.4	−153.5	−121.1	−130.1	−97.1	−94.3	−72.3	−55.7	−31.1	−28.2	−.01	−6.1

Note: Total $N = 600$. $\hat{c} - \hat{c}'$ refers to the point estimate of reduction in coefficients; FS refers to the Freedman and Schatzkin (1992) statistic. True effect is $\gamma_{01}^{(1)} - \gamma_{01}^{(4)}$ and estimated effect is $\gamma_{01}^{(1)} - \gamma_{01}^{(3)}$.

Table 4
Empirical Power for Sobel Statistic (When True Between-Group $\gamma_{02}^{(4)}$ Is .14)

Within-Group Coefficient $\gamma_{10}^{(4)}$		−.59	−.39	−.14	0	.14	.39	.59
Size	# Groups				CWC(M)			
5	120	.95	.96	.95	.95	.97	.96	.95
8	75	.80	.78	.80	.79	.79	.81	.79
12	50	.55	.54	.58	.56	.54	.60	.53
20	30	.22	.26	.24	.30	.25	.26	.23
					Grand-mean centering			
5	120	.99	1.00	.84	.50	.99	.99	.99
8	75	.93	.91	.84	.22	.91	.94	.92
12	50	.82	.79	.75	.09	.79	.83	.78
20	30	.60	.62	.53	.04	.58	.63	.57

Note: Empirical power is the proportion of Sobel statistics (absolute values) that are greater than 1.96, out of the 500 replications in a given condition. CWC(M) = centered within context with reintroduction of the subtracted means.

within the range of ± 8%). However, when the two coefficients differ, the absolute value of the percentage inaccuracy typically lies in the range of 70%–500%.

Empirical Power

Empirical power for testing the multilevel mediation effect (when the true between-group coefficient is small) is reported in Tables 4 and 5. Grand-mean centering is compared with CWC(M) in both tables. First, the CWC(M) panels in both tables show that the Freedman and Schatzkin statistic has greater power to detect mediation effects when there are a small number of groups with a large group size. Second, for both statistics, large numbers of groups with a small group size lead to more powerful tests. This is because the between-group coefficient becomes more influential when there are more groups, holding constant total sample size.

The grand-mean centering panels in Tables 4 and 5 need to be examined along with the first panels of confounding reported in Tables 2 and 3. For example, the first panel of Table 2 shows that when the within-group coefficient is −.59 and the between-group coefficient is .14, the Sobel statistic has a negative inaccuracy estimate of near −245%. This means that in Table 4, the corresponding power estimates were actually based on erroneous Sobel statistics. These seemingly high-power values actually correspond to an incorrect point estimate.

Empirical Type-I Error Rate

Tables 6 and 7 provide the Type-I error rate estimates for the two procedures. For the product-of-coefficients method in which a Sobel statistic is used, CWC(M) consistently showed a Type-I error rate less than .05 for all group sizes. The Type-I error rate for

Table 5
Empirical Power for F&S Statistic (When True Between-Group $\gamma_{02}^{(4)}$ Is .14)

Within-Group Coefficient $\gamma_{10}^{(4)}$		−.59	−.39	−.14	0	.14	.39	.59
Size	# Groups				CWC(M)			
5	120	.84	.95	.99	1.00	1.00	.96	.86
8	75	.78	.90	.98	.99	.98	.95	.84
12	50	.76	.85	.94	.96	.92	.86	.81
20	30	.70	.75	.84	.85	.87	.82	.71
					Grand-mean centering			
5	120	1.00	1.00	.84	.44	1.00	1.00	1.00
8	75	.98	.99	.89	.18	1.00	1.00	1.00
12	50	.94	.93	.86	.05	.98	1.00	.98
20	30	.77	.79	.64	.01	.95	.96	.94

Note: Empirical power is the proportion of the Freedman and Schatzkin (1992) statistics (absolute values) that are greater than 1.96, out of the 500 replications in a given condition. CWC(M) = centered within context with reintroduction of the subtracted means.

Table 6
Empirical Type-I Error Rate for Sobel Statistic
(When True Between-Group $\gamma_{02}^{(4)}$ Is 0)

Within-Group Coefficient $\gamma_{10}^{(4)}$		−.59	−.39	−.14	0	.14	.39	.59
Size	# Groups				CWC(M)			
5	120	.03	.03	.02	.02	.02	.02	.03
8	75	.04	.01	.02	.02	.03	.04	.02
12	50	.02	.01	.03	.02	.02	.02	.01
20	30	.01	.02	.01	.01	.01	.01	.01
					Grand-mean centering			
5	120	.99	.98	.99	.04	.97	.98	.99
8	75	.93	.94	.89	.03	.91	.93	.90
12	50	.79	.79	.79	.01	.76	.78	.80
20	30	.58	.61	.59	.00	.57	.66	.64

Note: Type-I error rates are the proportion of Sobel statistics (absolute values) that are greater than 1.96, out of the 500 replications in a given condition. CWC(M) = centered within context with reintroduction of the subtracted means.

grand-mean centering ranges from .57 to .99, when the within-group coefficient is different from zero. In other words, when the within-group effect has a coefficient of $\pm$.14, $\pm$.39, or $\pm$.59, the results show that grand-mean centering almost always produces false positives.

For the difference-in-coefficients method where Freedman and Schatzkin's (1992) statistic is used, the CWC(M) method of centering produced high Type-I error rates, ranging from .16 to .43. When grand-mean centering was used, similar patterns emerge as for the Sobel statistic. That is, when the within-group coefficient is not zero, the Type-I error rate is extremely high, ranging from .85 to 1.00.

Table 7
Empirical Type-I Error Rate for Freedman and Schatzkin (1992) Statistic (When True Between-Group $\gamma_{02}^{(4)}$ Is 0)

Within-Group Coefficient $\gamma_{10}^{(4)}$		−.59	−.39	−.14	0	.14	.39	.59
Size	# Groups				CWC(M)			
5	120	.40	.38	.19	.16	.19	.37	.41
8	75	.35	.32	.21	.19	.21	.34	.41
12	50	.40	.34	.22	.21	.22	.34	.43
20	30	.32	.32	.26	.17	.19	.31	.36
					Grand-mean centering			
5	120	1.00	1.00	1.00	.02	1.00	1.00	1.00
8	75	1.00	.99	1.00	.02	1.00	1.00	1.00
12	50	.97	.97	.99	.00	.98	.97	.96
20	30	.85	.90	.93	.00	.92	.91	.86

Note: Type-I error rates are the proportion of the Freedman and Schatzkin (1992) statistics (absolute values) that are greater than 1.96, out of the 500 replications in a given condition. CWC(M) = centered within context with reintroduction of the subtracted means.

Discrepancy between $\hat{c} - \hat{c}'$ and $\hat{a}\hat{b}$

Table 8 shows that under grand-mean centering, the average discrepancy between $\hat{c} - \hat{c}'$ and $\hat{a}\hat{b}$ is always zero, replicating the results of Krull and MacKinnon (1999). Under CWC(M), there are small discrepancies between $\hat{c} - \hat{c}'$ and $\hat{a}\hat{b}$ when the within-group coefficient is much lower than the between-group coefficient. This discrepancy is within the $\pm$ 5% range.

Discussion and Recommendations

This study shows that in multilevel mediation models such as the 2-1-1 model, the within-group effect in the 1-1 relationship can cause confounded estimates of multilevel mediation effects, which exist only at Level-2. In these models, Sobel statistics and Freedman and Schatzkin statistics based on grand-mean centering or no centering performed poorly because these statistics are based on a conflation of between-group and within-group effects.

Also, because grand-mean centering leads to high Type-I error rates, it can produce false positives when within-group effects are present but between-group mediation is absent. To interpret the damaging nature of the false positives, we continue with our example of the relationship among flexible work schedule (Level-2 antecedent), job satisfaction (Level-1 mediator), and employee performance (Level-1 outcome). Our results show that when the within-group relationship between job satisfaction and employee performance is strong and the between-group relationship between these two variables is weak, grand-mean-centering may lead researchers to incorrectly conclude that job satisfaction mediates the relationship between flexible work schedule and employee performance. In fact, the significant indirect effects under grand-mean-centering ($\hat{c} - \hat{c}'$ or $\hat{a}\hat{b}$) may be due solely to the strong within-

Table 8
Percentage Discrepancy Between $\hat{c} - \hat{c}'$ and $\hat{a}\hat{b}$ Under CWC(M) Versus Grand-Mean Centering

Within-Group Coefficient $\gamma_{10}^{(4)}$		−.59		−.39		−.14		0		.14		.39		.59	
		CWC(M)	Grand-mean	CWC(M)	Grand-mean	CWC(M)	Grand-mean	CWC(M)	Grand-mean	CWC(M)	Grand-mean	CWC(M)	Grand-mean	CWC(M)	Grand-mean
Size	# Groups	True between-group coefficient $\gamma_{02}^{(4)}$ is .14													
5	120	3.32	.00	−.44	.00	.86	.00	.11	.00	.00	.00	.15	.00	1.19	.00
8	75	−3.27	.00	.60	.00	.29	.00	−.27	.00	.01	.00	−2.53	.00	.12	.00
12	50	−.56	.00	−3.03	.00	.43	.00	−.78	.00	−.22	.00	−1.19	.00	−.06	.00
20	30	1.22	.00	.07	.00	.85	.00	.28	.00	−.31	.00	1.65	.00	−.26	.00
		True between-group coefficient $\gamma_{02}^{(4)}$ is .39													
5	120	.68	.00	−.37	.00	.21	.00	−.08	.00	.26	.00	.03	.00	.05	.00
8	75	.12	.00	1.68	.00	−.58	.00	.08	.00	−.38	.00	−.01	.00	.16	.00
12	50	−.92	.00	−1.27	.00	−.44	.00	−.59	.00	.07	.00	−.08	.00	.20	.00
20	30	2.05	.00	−.23	.00	.02	.00	−.18	.00	−.47	.00	−.04	.00	.05	.00
		True between-group coefficient $\gamma_{02}^{(4)}$ is .59													
5	120	.39	.00	.99	.00	−.05	.00	.22	.00	−.17	.00	.35	.00	.01	.00
8	75	.41	.00	1.13	.00	.53	.00	−.08	.00	.06	.00	.19	.00	.02	.00
12	50	1.05	.00	.79	.00	.16	.00	.47	.00	.56	.00	−.05	.00	.05	.00
20	30	−.47	.00	.13	.00	−1.19	.00	.44	.00	−.80	.00	.29	.00	−.01	.00

Note: Values reported are calculated as $\left((\hat{c} - \hat{c}') - \hat{a}\hat{b}\right) \times 100/\hat{a}\hat{b}$. CWC(M) = centered within context with reintroduction of the subtracted means.

group relationship, when this within-group relationship may be totally irrelevant for researchers who are interested in Level-2 mediation. Therefore, in this example and elsewhere, it is important to differentiate the multiple components of multilevel mediation effects. To accomplish this, between-group and within-group effects should be separated when testing multilevel mediation.

Potential Confounding in Testing 1-1-1 Models

Confounding similar to that in 2-1-1 models may exist for 1-1-1 models. In the 1-1-1 model, the patterns of confounded estimates could be more complicated. To provide a concrete example, assume that the following research question is asked: Will organizational identification mediate the effect of individuals' perceptions of transformational leadership on citizenship behavior? In this case, the antecedent, the mediator, and the outcome are all operationalized as Level-1 variables. To answer this research question based on nested data, we need to differentiate the between- and within-group relationships.

In contrast with 2-1-1 models where the mediation of interest could exist only across Level-2 units, researchers examining 1-1-1 models can investigate the within- or between-group mediation effects or both. That is, researchers may examine whether there is a group-level mediation effect, a within-group (frog-pond) mediation effect, or both types of effect. Does aggregated organizational identification mediate the influence of group-level transformational leadership on aggregated citizenship behavior? Does the relative standing of individuals on organizational identification mediate the relationship between the group-mean-centered transformational leadership perception on group-mean-centered citizenship behavior? Do the above two mediation effects differ in magnitude and, if so, which is stronger?

In 1-1-1 models, it might be inappropriate to label an effect at one level as a source of confounding, because researchers may be interested in both group-level and within-group-level effects. In either case, simply ignoring the two different sources of variation and estimating a ''combined'' 1-1-1 mediation effect will always be less informative than an examination that separately estimates both effects (i.e., using CWC(M)). To illustrate, compare Equations (16–18) with Equations (26–28) (see Table 1). In this first step, the total effect of X_{ij} on Y_{ij} (i.e., $\gamma_{10}^{(1)}$) is a combination of between-group ($\gamma_{01}^{(4)}$, from Equation [27]) and within-group effects ($\gamma_{10}^{(4)}$, from Equation [28]). In addition, the overall relationships between M_{ij} and Y_{ij} and between X_{ij} and M_{ij} can also be decomposed into between-group and within-group parts. Consequently, it is critical to look at the between-group mediation effect and within-group mediation effect separately in 1-1-1 models. This differentiation—possibly using CWC(M)—can provide a clearer view of the mediation mechanism at each level of analysis, including the possibility of completely equivalent effects across both levels. Combining the between-group and within-group parts into a single effect may lead to misrepresentations of mediation at Level 1, Level 2, or both.

Kenny, Korchmaros, and Bolger (2003) and Bauer et al. (2006) have discussed the complications in the 1-1-1 multilevel model. These authors offered methods to calculate the indirect effect, and its standard error, to account for the covariance between the Level-2 random effects. More recently, Pituch, Tate, and Murphy (in press) have discussed three-level

mediation models. However, none of these studies have explicitly partitioned within- and between-unit mediation effects at each level. Extending these prior studies, we propose partitioning and simultaneously examining the two mediation effects in 1-1-1 models, and recommend the use of CWC(M) as a technique for partitioning these two independent mediation effects.

Recommendations for Testing Multilevel Mediation Using CWC(M)

The central argument of this study is that some multilevel mediation effects need to be decomposed into Level-1 and Level-2 effects. This decomposition could provide clearer answers to research questions that focus on either Level-1 or Level-2 mediation relationships or both. Based on the results of our simulation study, we put forth a set of recommendations intended to assist researchers in choosing between various mediation methods, improving the accuracy of multilevel mediation testing, and increasing the consistency of result reporting.

First, it is important to note that strong multilevel theory is the basis for strong multilevel research design and mediation testing. As researchers often contend, mediation inference rests on the tripod of strong theory, sound research design and measurement, and appropriate statistical analysis (Mathieu & Taylor, 2007; Stone-Romero & Rosopa, 2008). Because mediation effects may not be distinguishable statistically from confounding and suppression effects (MacKinnon, Krull, & Lockwood, 2000), before choosing an analytical method to test mediation, researchers are urged to develop their mediation models based on solid multilevel theory and use valid multilevel research design and measurements to assess the multilevel phenomenon of interest.

After questions are developed based on sound theory, researchers must choose a model (e.g., a 2-1-1 model, a 1-1-1 model, or a 2-2-1 model) that can best answer the research questions and then design their studies accordingly. The potential confounding we discuss may not be relevant if researchers choose a 2-2-1 model. However, if a 2-1-1 or 1-1-1 model is used, researchers need to pay special attention to the differentiation of mediation effects at different levels.

Next, if using 2-1-1 or 1-1-1 models, researchers need to decide whether they are interested in only the between-group relationship, the within-group relationship, or both—although it is our belief that in the 2-1-1 model the focus may only be between groups. In each of the three situations, it is useful to differentiate the relationships at the two levels rather than combining them into a single estimate. In all cases, CWC(M) is recommended for formulating a multilevel mediation model. Table 1 provides the recommended procedures for conducting multilevel mediation testing for 2-1-1 and 1-1-1 models. In 1-1-1 models, the antecedent, the mediator, and the outcome variables are all group-mean centered, with group means entered at Level 2.

Finally, after analyzing data based on CWC(M), we believe researchers should report results at both levels of analysis, regardless of the level at which the effect should theoretically exist. Reporting both the Level-1 and Level-2 coefficients and mediation effects would facilitate the comparison between levels. This makes it easier for future meta-analytic research to accumulate and synthesize study results, as well as provide useful information for researchers curious about differences in effects across levels of analysis.

Study Limitations

This study has several limitations. This study, as with all simulation studies, offers results that are subject to the influence of the parameter starting values, model assumptions, and estimation methods. Using an alternative set of assumptions, starting values, and estimation methods may change the results accordingly. In addition to these general limitations, there are several limitations more specific to this study. First, the total sample size was fixed at 600 in the simulation and thus the independent effects of group size and the number of groups were not examined. However, given the practical limitations involved in multilevel data collection, we believe this total sample size and the trade-off between group size and the number of groups reflect the real-life situation that researchers often face in field studies.

Related to this limitation, only four group sizes were investigated here. Although more group size categories would better illustrate the influence of group size, we believe the current four sizes already show a clear pattern of influence by group size, and this pattern may not change as a function of more group sizes.

Next, the between- and within-group coefficients were set in a way that has absolute values of 0, .14, .39, and .59. Although larger coefficients could have been used (e.g., .90 and 1.20) in addition to these values, again, we believe that adding coefficients of large magnitude to the current list would show a similar pattern of results to those illustrated above.

Third, it is noteworthy that we did not include a random slope in our multilevel models. We made this choice for two reasons. First, our interest was not in estimating random slopes—although we foresee no difficulty in generalizing our findings to models containing random slopes. Second, adding random slopes to our models has the potential to increase rates of nonconvergence, so we excluded them.

Finally, because of its focus on multilevel mediation, this study does not explicitly discuss moderation or moderated mediation in multilevel settings. Whereas mediation analysis involves determining the extent to which an intervening variable transmits the effect of a predictor to a criterion, moderation analysis involves determining the extent to which a moderator variable affects the relationship between a predictor and a criterion. These two kinds of effect are often confused (Baron & Kenny, 1986), but are conceptually distinct. In the multilevel setting that was the focus of this article, we recommend estimating separate between-group and within-group effects to avoid confounding, but this idea is conceptually distinct from claiming that any of the effects involved (direct or indirect) are moderated. The between- and within-effects are distinct effects linking constructs with different interpretations at the between- and within-levels. That said, there do exist ways to combine mediation and moderation (e.g., Bauer et al., 2006; Edwards & Lambert, 2007), but extending those methods to accommodate separation of between- and within-effects lies outside the scope of this study. Although the lack of discussion of moderated multilevel mediation models is a limitation of this study, we believe that the general principle of separating between-group and within-group mediation effects and the specific treatment using CWC(M) in theory can be readily extended to the more complicated moderated mediation models. For example, in models such as 2 × (2-1-1) and 2 × (1-1-1), the Level-2 moderator can be allowed to have different moderating effects at the between-group level versus the within-group level using

CWC(M). Exploring the complications and specific decision points in moderated multilevel mediation models would be a fruitful avenue for future research.

Future research may also investigate longer and/or multiple mediation chains in multilevel settings. As Taylor, MacKinnon, and Tein (2008) show, bias-corrected bootstrapping may be the most accurate method to estimate three-path mediation effects in a single-level setting. There remain questions with regard to how multiple or longer chained multilevel mediation hypotheses could be tested, but regardless of the length of mediation chains and the presence or absence of moderators, it is important to consider the possibility of different effects within versus between groups.

In conclusion, we have shown that tests of multilevel mediation can be problematic when between-group variation in a Level-1 variable is not explicitly separated in a test of 2-1-1 mediation—the same would be true for 1-1-1 mediation tests. By adhering to traditional recommendations for testing mediation with multilevel data, researchers may be making one hypothesis (i.e., group-level mediation) while testing another (i.e., mediation that conflates between-group and within-group effects). By keeping an eye on the within-group effect in a multilevel mediation model, researchers not only may gain insight into differences in effect across levels of analysis, but also may increase the precision of their tests of multilevel mediation.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173-1182.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, *11*, 142-163.

Busing, F. (1993). Distribution characteristics of variance estimates in two-level models. Unpublished manuscript.

Chen, G., & Bliese, P. D. (2002). The role of different levels of leadership in predicting self- and collective efficacy: Evidence for discontinuity. *The Journal of Applied Psychology*, *87*, 549-556.

Chen, G., Kirkman, B.L., Kanfer, R., Allen, D., & Rosen, B. (2007). A multilevel study of leadership, empowerment, and performance in teams. *The Journal of Applied Psychology*, *92*, 331-346.

Cheung, G. W., & Lau, R. S. (2008). Testing mediation and suppression effects of latent variables: Bootstrapping with structural equation models. *Organizational Research Methods*, *11*, 296-325.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Dansereau, F., Alutto, J. A., & Yammarino, F. J. (1984). *Theory testing in organizational behavior: The variant approach*. Englewood Cliffs, NJ: Prentice Hall.

Davis, J. A., Spaeth, J. L., & Huson, C. (1961). A technique for analyzing the effects of group composition. *American Sociological Review*, *26*, 215-225.

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, *12*, 1-22.

Freedman, L. S., & Schatzkin, A. (1992). Sample size for studying intermediate endpoints within intervention trials of observational studies. *American Journal of Epidemiology*, *136*, 1148-1159.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, *24*, 623-641.

Hofmann, D. A., Griffin, M. A., & Gavin, M. B. (2000). The application of hierarchical linear modeling to organizational research. In K. J. Klein & S. W. J. Kozlowski (eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 467-511). San Francisco, CA: Jossey-Bass.

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, *8*, 115-128.

Klein, K. J., & Kozlowski, S. W. J. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods*, *3*, 211-236.

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

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group based intervention studies. *Evaluation Review*, *33*, 418-444.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, *36*, 249-277.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Taylor & Francis.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science, 1,* 173-181.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test the significance of the mediated effect. *Psychological Methods*, *7*, 83-104.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*, 99-128.

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*, 41-62.

Mathieu, J. E., DeShon, R. P., & Bergh, D. D. (2008). Mediational inferences in organizational research: Then, now and beyond. *Organizational Research Methods*, *11*, 203-223.

Mathieu, J. E., & Taylor, S. R. (2007). A framework for testing meso-mediational relationships in organizational behavior. *Journal of Organizational Behavior*, *28*, 141-172.

Muller, D., Yzerbyt, V. Y., & Judd, C. M. (2008). Adjusting for a mediator in models with two crossed treatment variables. *Organizational Research Methods*, *11*, 224-240.

Ostroff, C. (1993). Comparing correlations based on individual level and aggregate data. *The Journal of Applied Psychology*, *78*, 569-582.

Pituch, K. A., & Stapleton, L. M. (2008). The performance of methods to test upper-level mediation in the presence of nonnormal data. *Multivariate Behavioral Research*, *43*, 237-267.

Pituch, K. A., Tate, R. L., & Murphy, D. L. (in press). Three-level models for indirect effects in school- and class-randomized experiments in education. *Journal of Experimental Education*.

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

Seibert, S. E., Silver, S. R., & Randolph, W. A. (2004). Taking empowerment to the next level: A multiple-level model of empowerment, performance, and satisfaction. *Academy of Management Journal*, *47*, 332-349.

Snijders, T. A., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural models. In S. Leinhardt (ed.), *Sociological Methodology 1982* (pp. 290-312). San Francisco, CA: Jossey-Bass.

Stone-Romero, E. F., & Rosopa, P. J. (2008). The relative validity of inferences about mediation as a function of research design characteristics. *Organizational Research Methods*, *11*, 326-352.

Taylor, A. B., MacKinnon, D. P., & Tein, J.Y. (2008). Tests of the three-path mediated effect. *Organizational Research Methods*, *11*, 241-269.

Zhen Zhang is an assistant professor of management at Arizona State University. He received his PhD in human resources and industrial relations from the University of Minnesota. His research focuses on leadership process and development, the interfaces between organizational behavior and entrepreneurship, and research methods.

Michael J. Zyphur received his PhD in industrial and organizational psychology from Tulane University. His research interests include research methods and the biological basis of organizational behavior. He is currently an assistant professor at the University of Washington, Bothell.

Kristopher J. Preacher is an assistant professor of quantitative psychology at the University of Kansas. His research focuses on the use of latent variable analysis and multilevel modeling to analyze longitudinal and correlational data. Other interests include developing techniques to address mediation, moderation, and model evaluation and selection.