Predicting group-level outcome variables: An empirical comparison of analysis strategies

Lynn Foster-Johnson¹ · Jeffrey D. Kromrey²

Published online: 5 March 2018
© Psychonomic Society, Inc. 2018

Abstract

This study provides a review of two methods for analyzing multilevel data with group-level outcome variables and compares them in a simulation study. The analytical methods included an unadjusted ordinary least squares (OLS) analysis of group means and a two-step adjustment of the group means suggested by Croon and van Veldhoven (2007). The Type I error control, power, bias, standard errors, and RMSE in parameter estimates were compared across design conditions that included manipulations of number of predictor variables, level of correlation between predictors, level of intraclass correlation, predictor reliability, effect size, and sample size. The results suggested that an OLS analysis of the group means, with White's heteroscedasticity adjustment, provided more power for tests of group-level predictors, but less power for tests of individual-level predictors. Furthermore, this simple analysis avoided the extreme bias in parameter estimates and inadmissible solutions that were encountered with other strategies. These results were interpreted in terms of recommended analytical methods for applied researchers.

Keywords Micro–macro data · Group-level outcomes · Analysis of group means

Analyzing multilevel data

An increasing number of investigations have examined methods for correctly analyzing data that are collected in multilevel contexts. Multilevel data structures occur in almost every discipline. However, the degree to which various disciplines acknowledge the complexities of these data and concomitantly seek to analyze them appropriately varies. Considerable work in education, psychology, medicine and management has acknowledged the intricacies of multilevel data and recommended sophisticated analysis methods to address the challenges. Other fields have seen less of an emphasis on capturing the complexities of these data and developing appropriate analysis methods.

Multilevel data structures are data that naturally occur in a hierarchically ordered system (see Hofmann, 1997). Common multilevel data structures in education include students nested within classrooms, schools, and districts. In management, data structures are often centered on employees working within teams, units, or departments (see Wood, Van Veldhoven, Croon, & de Menezes, 2012). In clinical medicine, patients are often clustered within clinical trial sites or teams of attending physicians. Medical services data structures may find individual patient services embedded within hospital service areas or referral regions (see Fisher, Bynum, & Skinner, 2009).

A substantial body of work has been developed for analyzing multilevel data in which the outcome variable is measured at the individual level (see, e.g., Raudenbush & Bryk, 2002). Often referred to as the *macro–micro* data situation (see Snijders & Bosker, 2012), the dependent variable Y is measured at the lower level (e.g., individual) and is assumed to be affected by explanatory variable(s) X , which are also measured at the lower level, and group-level variables (Z), which are measured at a higher level. In education and social sciences, the most common analysis method used for these data structures is hierarchical linear modeling (Raudenbush & Bryk, 2002) or random-effects models (Hedeker, Gibbons, & Flay, 1994).

Less work has been devoted to the *micro–macro* data situation, in which Y is measured at the higher (group) level, and

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13428-018-1025-8>) contains supplementary material, which is available to authorized users.

✉ Lynn Foster-Johnson
Lynn.Foster-Johnson@Dartmouth.edu

¹ Geisel School of Medicine, Dartmouth College, Hanover, NH, USA

² College of Education, University of South Florida, Tampa, FL, USA

corresponding explanatory variables are measured at the individual level (X) and at the group level (Z). Generally, there have been two approaches to analyzing micro–macro data. Although this method is not generally accepted, one could analyze the data at the individual level, essentially disaggregating the group-level data and repeating the group variable scores for each individual in the group. Such analyses usually yield biased estimates of the standard errors and grotesquely inflated Type I error rates for hypothesis tests. The more popular analysis method for micro–macro data analysis is to aggregate the data measured at a lower level (i.e., individual) to a higher level—generally the level at which Y , the dependent variable, is measured. Under these data conditions, the level at which Y is measured is often a naturally occurring group, such as a team, a classroom, a hospital ward, or a department. In this analysis approach, the group means of the explanatory variables are used as scores on variables in the subsequent analyses conducted at the group level.

Although an aggregated analysis is a simple approach to conduct a micro–macro analysis, a number of researchers have expressed concern about the suitability of aggregated data analysis methods for multilevel types of data. In addition to the loss of information at the individual level, the reduction of variability in the data due to aggregation leads to inaccurate estimates of standard errors and bias in regression parameters (Clark & Avery, 1976; Richter & Brorsen, 2006).

Recent work by Myer, Thoroughgood, and Mohammed (2016) provides an example of data collected using a micro–macro structure. Myer and his colleagues examined whether being ethical comes at a cost to profits in customer-oriented firms, by looking the interaction between service and ethical climates on company-level financial performance. Their study used a sample of 16,862 medical sales representatives spread across 77 subsidiary companies of a large multinational corporation in the health care product industry. Climate variable predictors were the data collected from an annual employee attitude survey, and the outcome variable was subsidiary company-level financial performance. The individual-level climate data were aggregated by subsidiary company for analysis. Using hierarchical multiple regression, they found a significant interaction that accounted for an additional 6% of the variance in financial performance beyond the control and main effects. An analysis of the simple slopes showed that service climate was positively related to financial performance when the ethical climate was also high, but not when it was low.

Aim of present study

The popularity of multilevel data requires an understanding of appropriate analytic methods for micro–macro data. Our aims of this study are twofold:

- To present an alternative approach for analyzing micro–macro data proposed by Croon and van Veldhoven (2007) that uses adjusted group means.
- To compare the unadjusted and adjusted-group-means approaches using simulated data. Type I error, power, bias, RMSE, and model convergence will be examined.

The article is organized as follows: First we describe the adjusted-group-means analysis approach proposed by Croon and van Veldhoven (2007) and provide formulaic treatment of this method. We then provide the results of a simulation study comparing the typical method for analyzing micro–macro data (unadjusted group means) with the adjusted-group-means approach.

Conducting a micro–macro analysis

Recent work in latent trait modeling has suggested that group-level effects might be better modeled by treating the macro-level units as the unit of analysis and the micro-level data as indicators. This is similar to a standard structural equation modeling context in which the latent variables are measured using indicator variables on which all subjects have scores. In an extension of this approach to multilevel data, the person-as-variables approach uses the subjects as indicators for the unobserved score at the group level. Multiple individual-level indicators define a latent construct at both the individual and group levels. The analysis is treated as a restricted confirmatory factor analysis (CFA) in which factor loadings are fixed, rather than estimated, and person-specific data are used for modeling means as well as covariances (Bauer, 2003; Curran, 2003; Mehta & Neale, 2005).

Borrowing from the persons-as-variables approach, Croon and van Veldhoven (2007) presented a formalized representation of an alternative aggregation strategy, suggesting that a simple substitution of group means for predictors measured at the individual level leads to biased estimates of the regression parameters. They proposed, instead, a method that uses adjusted group means for the individual-level predictors followed by an ordinary least squares (OLS) analysis. Similar to what one would find if one used the lower-level units as indicators for a latent variable, this adjustment takes into account all of the observed scores on the individual (X) and group-level (Z) explanatory variables in each group for the calculation of the adjusted group mean value. We provide an overview of this approach in the next section and refer the reader to Croon and van Veldhoven for more details.

Croon and van Veldhoven (2007) provided results of a small simulation demonstrating their approach in which several design factors were manipulated, including the number of groups, group size, intraclass correlation (ICC), and correlation between X (individual-level measures) and Z (group-level

measures) variables. The simulation results suggested that in the unadjusted approach, two design factors led to severe downward bias of the slopes at the individual-level regression: group size and the size of the intraclass correlation. Bias for unadjusted parameter estimates was smaller for larger groups and for higher values of ICC. The effects of the size of the correlation and the number of groups on bias were much smaller. At the group level, the bias in the estimation of the regression coefficient was less extreme. Both the size of the correlation between X and Z and the smaller group size were associated with greater bias. The adjusted regression analysis reduced the bias in the parameter estimates. Over all the conditions, the percentage of conditions with bias in estimating the X and Z predictors was significantly reduced, and no systematic patterns in bias due to the design effects was found.

In a second analysis, Croon and van Veldhoven (2007) compared the unadjusted and adjusted results of a regression analysis of financial performance of business units on four psychological climate scales measured at the individual-level (employee). They found in the unadjusted analysis that three of the individual-level explanatory variables were significant predictors of financial performance, whereas the adjusted analysis resulted in only one explanatory variable reaching statistical significance. The parameter estimates and standard errors from the adjusted analysis were larger than the corresponding coefficients from the unadjusted analysis. They attribute this to the adjustment in their procedure that transforms the measurement scale of the individual-level explanatory variables when no group explanatory variables are involved in the analysis. Robust standard errors using the White–Davidson–MacKinnon correction procedure, were the same or less than the OLS standard errors, but did not substantively change the results.

Although not explicitly presented by Croon, the typical standard error for a partial regression slope is implied:

$$SE_i = \sqrt{\frac{MS_{\text{residual}}}{(SS_i)(1-R_{i,\text{other predictors}}^2)}} \quad (1.1)$$

where SE_i is the standard error of the i th predictor, MS_{resid} is the mean square residual from the regression model, SS_i is the sum of squares of the i th predictor, and $(1-R_{i,\text{other predictors}}^2)$ is the tolerance of the i th predictor.

The latent-variable approach

Croon and van Veldhoven (2007) presented a latent variable approach to the analysis of individual- and group-explanatory

variables in predicting a group outcome variable Y . Given a set of linear equations in which the relationship between the group scores on explanatory variables Z (observed) and ξ (latent) and the outcome variable Y is:

$$y_g = \beta_0 + \beta_1 \xi_g + \beta_2 z_g + \epsilon_g \quad (1.2)$$

The latent group-level variable ξ represents the unobserved variable that gives rise to the observed individual-level explanatory variable X . Each individual's score on X , x_{ig} , is treated as an indicator for the unobserved group score (ξ_g). The unobserved group-level score ξ_g may be correlated with the observed group-level variable Z , and both may have an effect on the group-level outcome variable Y . The error component ϵ is assumed to be homoscedastic, or to have a constant variance for all groups.

All three parameters in Eq. 1.2 are defined at the group level, but ξ_g is not an observed variable. The relationship between ξ_g and x_{ig} must be modeled as: $x_{ig} = \xi_g + \nu_{ig}$

where the variance of ξ_g is denoted by σ_ξ^2 and the variance of the disturbance term ν_{ig} , by σ_ν^2 . The within-group variance σ_ν^2 , is assumed to be constant for all subjects and groups. The total variance of X , σ_X^2 , is modeled as $\sigma_X^2 = \sigma_\xi^2 + \sigma_\nu^2$.

Given a multilevel data configuration in which the variables are observed rather than latent, the typical method of analysis would be as follows, where y_g is the score of group g on a group-level outcome variable Y , and $\bar{x}_g$ is individual-level variable(s) aggregated to the group level, and z_g is a group-level explanatory variable(s).

$$E(y_g | \bar{x}_g, z_g) = \beta_0 + \beta_1 \bar{x}_g + \beta_2 z_g \quad (1.3)$$

This equation expresses the relationship between the group outcome variable (y_g) and two observed quantities: group means $\bar{x}_g$ and z_g . The aggregated analysis solution depicted in Eq. 1.3 would be appropriate if it yielded results that are the same as those values derived from the model depicted in Eq. 1.2. However, the variance of the variable $\bar{x}_g$ in Eq. 1.3 is $\sigma_{\bar{x}}^2 = \sigma_\xi^2 + \sigma_\nu^2/n_g$ rather than σ_ξ^2 from the previous section, and the regression coefficients and standard errors will typically differ in the two equations. This bias results from treating the latent variable ξ as if it were observed.

Croon and van Veldhoven (2007) derived the regression equation relating the observed variables X and Z to Y while avoiding the bias presented in Eq. 1.3:

$$E(y_g | \bar{x}_g, z_g) = \beta_0 + \beta_1 \left[(1-w_{g1})\mu_\xi + w_{g2}(z_g - \mu_Z) + w_{g1}\bar{x}_g \right] + \beta_2 z_g \quad (1.4)$$

The bracketed expression following β_1 is the adjusted mean that must replace $\bar{x}_g$ in Eq. 1.3. The values w_{g1} and w_{g2} are weights applied to the observed data, providing the adjusted

means. With a single X variable and a single Z variable, these weights are obtained as:

$$w_{g1} = \frac{\sigma_{\xi}^2 \sigma_z^2 - \sigma_{\xi z}^2}{(\sigma_{\xi}^2 + \sigma_{\nu}^2/n_g) \sigma_z^2 - \sigma_{\xi z}^2} \quad (1.5)$$

$$w_{g2} = \frac{\sigma_{\xi z} \sigma_{\nu}^2/n_g}{(\sigma_{\xi}^2 + \sigma_{\nu}^2/n_g) \sigma_z^2 - \sigma_{\xi z}^2} \quad (1.6)$$

The use of these weights in the bracketed expression in 1.4 “shrinks” the group mean $\bar{x}_g$ toward the estimated population mean μ_{ξ} (i.e., removing the excess variability in $\bar{x}_g$) and also adjusts for the deviation of Z_g from the estimated population mean of Z . As is indicated in Eq. 1.6, the latter adjustment only occurs if covariance is present between ξ and Z . With multiple predictors at either level, the scalar quantities are replaced with the analogous covariance matrices. The resulting weights are used to obtain the adjusted group means on the X variables:

$$\tilde{x}_g = (1-w_{g1})\mu_{\xi} + w_{g2}(z_g - \mu_Z) + w_{g1}\bar{x}_g \quad (1.7)$$

To obtain unbiased estimates of the true regression coefficients, β_j , one must regress scores y_g on the adjusted group means $\tilde{x}_g$ and z_g rather than the group means $\bar{x}_g$. The adjusted group mean is the expected value of the unobserved variable ξ_g taking into account all of the observed individual- (X) and group- (Z) level explanatory variables in group g . Because of the exchangeability of the individuals within the group, their scores have constant weights in the expression for the best linear unbiased predictor, which implies that the group mean $\bar{x}_g$ is sufficient for the prediction of ξ_g . Specifics of the CV approach can be found in Croon and van Veldhoven (2007, especially pp. 51 to 52).

Considerations in micro–macro modeling

The analytical approach suggested by Croon and van Veldhoven (2007) may provide a viable strategy that is superior to simply aggregating the individual-level (X) variables to the higher, group level (Z) and conducting an OLS analysis on the resulting group-level data. However, the simulation presented in their article was limited in scope and did not explore the performance of their approach across broader and more realistic research conditions. As such, before their analytic recommendation can be completely supported, a variety of data conditions and considerations must be taken into account.

Sample size (number of groups) and group size (number within group) Sample size and the number of observations within each group are analysis issues that have an effect on

results when data are aggregated from the individual data to the group level.

Sample size considerations include ensuring sufficient numbers of groups to achieve statistical power and reasonable external validity of the findings. When conducting a micro–macro analysis using the averages of the individual-level variables at the group level, the sample size becomes the number of higher-level groups. As with a nonmultilevel analysis, the larger the number of groups, the greater statistical power and precision of analyses (Barcikowski, 1981; Hopkins, 1982).

In the multilevel context, most of the simulation studies find minimal bias in the estimates of fixed effects that can be attributed to sample size or number of groups (Clarke & Wheaton, 2007; Maas & Hox, 2005; Newsom & Nishishiba, 2002). However, sample- and group-sizes appear to have a greater impact on the estimation of random effects in multilevel models. Smaller sample- and group-sizes have been linked to greater bias in the random effects parameter estimates for both the intercept and the slope (Clarke & Wheaton, 2007; Maas & Hox, 2005; Mok, 1995). Biased variance estimates have been reported with designs having as few as two to five groups (Clarke, 2008; Mok, 1995), and some studies report biased estimates with sample sizes as large as 30 groups. With most studies, as sample sizes approach 50, bias in the variance estimates is eliminated. Some studies, however, are suggesting that this issue is more complex. Researchers are now finding that a combination of sample- and group-sizes yields the best information for reducing bias in multilevel random effects (Snijders & Bosker, 2012). For example, Clarke and Wheaton (2007) noted that a minimum of 100 groups with at least ten observations per group are necessary to eliminate bias in the intercept variance and for the slope variance, a minimum of 20 observations per group for at least 200 groups are necessary. Bias was more evident in multilevel models with singleton groups (groups with only one observation). In contrast, Bell-Ellison, Ferron, and Kromrey (2008) found low levels of bias for all parameter estimates in their investigation of sparse data structures in multilevel models. Singletons had no notable effect on bias with large numbers of groups, and only a small effect with fewer groups. Similar results were found for Type I error and statistical power.

Reliability of regressors Measurement error and the reliability of regressors, whether at the individual (X) or the group (Z) level are important for both macro and micro research approaches. If measurement errors affect the response variable only, then few difficulties are encountered as long as the measurement errors are uncorrelated random variables with zero mean and constant variance. These errors are incorporated into the model error term. However, when the measurement errors affect the regressor variables, the situation becomes much more complex.

The true value of the regressor is comprised of the observed value and the measurement error with an expected mean value of zero and a constant variance. This error must be modeled in the regression equation along with the standard error term associated with the response variable. Applying a standard least squares method to the data (and ignoring the measurement error) produces estimators of the model that are no longer unbiased. Unless there are no measurement errors in the regressors, $\hat{\beta}_1$ is always a biased estimator of β_1 (Cochran, 1968; Davies & Hutton, 1975). The detrimental effect of bias has been demonstrated with other multivariate statistical techniques, such as discriminant function analysis (see Kromrey, Yi, & Foster-Johnson, 1997), and canonical correlation (Thompson, 1990, 1991). Reliability of regressors also impacts Type I error and statistical power in regression. Kromrey and Foster-Johnson (1999b) found that with perfectly reliable regressors ($r_{xx} = 1.0$), error control was maintained. However, regressors with reliabilities of .80 rapidly resulted in elevated levels of Type I errors, particularly with models containing ten regressors. Type I errors increased as reliability of regressors decreased. In addition, lower reliability of regressors was associated with substantially lower levels of power.

The often quoted “gold standard” for acceptable levels of regressor internal consistency is $\alpha = .70$ or above (Nunnally, 1978). Pedhazuer and Schmelkin (1991), however, have suggested that the more important reliability consideration has to do with the type of decisions and the possible consequences of those decisions, rather than an absolute reliability value. Hence, for early stages of research, relatively low reliabilities are tolerable, whereas greater levels of reliability are needed when measures are used to determine differences among groups, and very high reliabilities are needed when scores are used for making important decisions about individuals. These suggestions may provide some guidance about measurement reliability with the more complex data configurations associated with micro–macro and macro–micro analysis approaches.

In the multilevel or macro–micro data context, reliability of regressors, whether at the individual (X) or the group (Z) level is also important. For a number of years controversy has surrounded the meaning of measures in the multilevel context when aggregation occurs. Referred to as *isomorphism*, the presumption is that there is a one-to-one correspondence between measures, even though they occur at different levels. Numerous scholars have noted that isomorphism cannot be automatically assumed in cases of multilevel data, thus drawing into question the accuracy of internal consistency and measurement precision claims (see Bliese, 2000; Bliese, Chan, & Ployhart, 2007; Chan, 1998, 2005; Kozlowski & Klein, 2000; Mossholder & Bedeian, 1983; O’Brien, 1990; Snijders & Bosker, 2012; Van Mierlo, Vermunt, & Rutte,

2009). In addition, a growing number of researchers have reported that measurement errors in multilevel models can bias fixed- and random-effects estimates, and have suggested methods for specifying and adjusting for the measurement error (Hutchison, 2003; Huynh, 2006; Longford, 1993; Raykov & Penev, 2009; Woodhouse, Yang, Goldstein, & Rasbash, 1996). However, response to these recommendations has been limited. In their review of methodological issues in multilevel modeling, Dedrick and his colleagues noted that only 18% of the 99 studies in their review reported the potential impact of measurement error on the resulting models (Dedrick et al., 2009). In an analysis of the effect of regressor reliability in multilevel models, Kromrey and his colleagues (2007) found that model convergence improved as the reliability of the regressors increased. With perfect regressor reliability, no bias was detected in the fixed or random effects. When regressor reliability was less than 1.00, statistical bias was positive for random effects and negative for fixed effects. Similar effects due to reliability of the regressors were seen for Type I error control and statistical power (Kromrey et al., 2007).

Degree of clustering Inherent in multilevel data analysis is that the data are collected at different levels, representing the clustering that is evident in naturally occurring hierarchies. There are numerous ways for determining the degree to which clustering exists in these data configurations. WABA, r_{wg} , and intraclass correlations (ICC) are the most common approaches used to justify a multilevel analysis or aggregation. We focus on ICC in more detail.

ICC is a measure of the degree of clustering that is due to the unit or naturally occurring hierarchy. A major issue with clustered data is that the observations within a cluster are not independent. Ignoring intra-cluster correlations could lead to incorrect standard errors, confidence intervals that are too small, and biased parameter estimates and effect sizes. Several versions of ICC have been developed (see Shrout & Fleiss, 1979). One of the most popular variations on ICC in the multilevel context is based on a one-way random effects analysis of variance (see Raudenbush & Bryk, 2002).

In most situations, the numeric value of ICC tends to be small and positive. Several authors have provided guidelines for interpreting the magnitude of ICCs with small, medium, and large values of ICC reported as .05, .10, and .15 (Hedges & Hedberg, 2007; Hox, 2002). The cluster effect, however, is a combination of the ICC and the cluster size. Small ICCs combined with large cluster sizes can still affect the validity of statistical analyses. Maas and Hox (2005) report that the largest bias for parameters estimates (both fixed and random) was found in conditions with the smallest sample sizes in combination with the highest ICC.

Correlation between regressors Correlation between regressors measured at the individual level (X s) has the same impact

in the multilevel context as in single-level data configuration. Similarly, the correlation between the variables measured at the individual level (X) and those measured at the group level (Z) also affects the outcomes. In the macro-micro context, it is well-known that severe collinearity presents problems in multiple regression analysis (see Cohen & Cohen, 1983; Kromrey & Foster-Johnson, 1999b). As collinearity between X s increases, Type I error rates increase and power decreases (see Kromrey & Foster-Johnson, 1998, 1999a, 1999b). In the multilevel context, high regressor intercorrelations and cross-level correlations have been found to be associated with model nonconvergence, greater statistical bias for the parameter estimates, increased Type I rates, and lower statistical power (Kromrey et al., 2007).

Homoscedasticity and normal distribution of the residuals

The assumption of homoscedasticity underlying regression in the micro-macro context is the same as that for the macro-micro multilevel data structure. Violations of these assumptions, when they are moderate, do not result in inaccurate parameter estimates or standard errors—especially if the sample size is not too small. There are several statistical methods for correcting heteroscedasticity if the violations become more severe. The most popular approach is known as heteroscedasticity consistent covariance matrix (HCCM) and is based on the work of White (1980). In his article, White presented the asymptotically justified form of the HCCM, which is generally referred to as HC0. Because of concerns about the performance of HC0 in small samples, MacKinnon and White (1985) developed three alternative estimators, known as HC1, HC2, and HC3, which were expected to have superior properties in small samples. Simulation studies comparing the performance of the correction approaches generally suggest that HC0 is biased downward for small sample sizes (Cribari-Neto, Ferrari, & Cordeiro, 2000; Cribari-Neto & Zarkos, 2001) and that HC3 provides better performance (Cai & Hayes, 2008; Cribari-Neto, Ferrari, & Oliveira, 2005; Long & Ervin, 2000).

In the multilevel context, Maas and Hox (2004) compared the standard errors from multilevel analysis and robust standard errors on the group-level parameter estimates of a multilevel regression and found that nonnormal residual errors at this level had little or no effect on the estimates of the fixed effects. The estimates of the regression coefficients were unbiased and both the multilevel and robust standard errors were accurate. However, nonnormally distributed residuals at the group level did have an effect on the parameter estimates of the random part of the model. Although the estimates were unbiased, the standard errors were not always accurate and the robust errors tended to perform better than the multilevel standard errors. If the distribution of the residuals was symmetric, the robust standard errors worked relatively well, but when the group-level residuals were skewed, neither the multilevel nor

the robust standard errors could compensate unless the number of groups was at least 100.

Simulation study

The primary purpose of the simulation study was to expand the scope of Croon and van Veldhoven (2007) by comparing the performance of their recommended approach with the traditional group aggregation analysis across broader and more realistic research conditions. In this context, we wanted to confirm their statistical bias results, provide Type I error and statistical power estimates, and test the viability of the less computationally complex alternative of aggregating on group means. A comparative investigation that is specific to data instances in which the dependent variable is measured at the group level is an important extension of the work of Croon and van Veldhoven and may provide an alternative to the typical group means aggregation approach currently used with data configurations such as these.

Method

The statistical performance of the Croon method (with [CV-W] and without White's, 1980, adjustment [CV]), and a traditional regression analysis of group means using sample means of individual-level predictors (with [GRP-W] and without White's adjustment [GRP]), was investigated using Monte Carlo methods, in which random samples were generated under known and controlled population conditions. We assumed that the individual-level measures would be indicators of the group-level construct, in which the scores associated with individuals in a group are interchangeable.

Our simulation was based on the following full-model equation:

$$y_g = \beta_0 + \bar{X}_g^T \beta_X + Z_g^T \beta_Z + \varepsilon_g \quad (1.8)$$

where $\bar{X}_g^T$ and Z_g^T are the row vectors of population mean true scores for the individual-level variables and the population true scores for the group-level variables, respectively, for group g , β_X and β_Z are the column vectors of partial regression slopes for the individual- and group-level variables, respectively, and ε_g is the residual term for group g .

Conditions

The Monte Carlo study included ten factors in the design: the number of individual- and group-level regressors; the correlation between the individual- and group-level regressors; cross-level correlations; reliability of regressors; the effect size; the

intraclass correlations; the number of groups; and the sample size in each group.

Number of regressors We varied the number of regressors at the individual (X) and at the group (Z) level. At the individual level (P_X), we included models with three, five, and seven individual-level regressors, extending the number of regressors from what was tested by Croon and Van Veldhoven (2007) to models that are more typical of the data analyzed by applied researchers. At the group level (P_Z), we included models with two and four group-level regressors.

Correlation between individual- (R_X) and group- (R_Z) level regressors We varied the correlation between the individual regressors by levels that would be considered low, medium, and high interregressor correlations ($\rho_X = .10, .30$, and $.50$). Correlations between group-level regressors were varied by values of ($\rho_Z = .20, .40$, and $.60$).

Cross-level correlations Cross-level correlations were established as correlations of group means of individual-level predictors and values of group-level predictors (i.e., the level-2 component of level-1 regressors or their cluster mean). Cross-level correlations (R_XZ) were set to zero, moderate, and high correlations ($\rho_{XZ} = 0, .30$, and $.50$). These values allowed comparison to Croon and van Veldhoven (2007) as well as providing performance information on a scenario in which cross-level correlations were high.

Reliability of regressors (R_XX) Measurement error was simulated in the data (following the procedures used by Maxwell, Delaney, & Dill, 1984; and by Jaccard & Wan, 1995), by generating two normally distributed random variables for each regressor (one to represent the “true scores” on the regressor, and one to represent measurement error). Fallible observed scores on the regressors were calculated (under classical measurement theory) as the sum of the true and error components. The reliabilities of the regressors were controlled by adjusting the error variance relative to the true score variance

$$\rho_{xx} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

where σ_T^2 and σ_E^2 are the true and error variance, respectively, and ρ_{xx} is the reliability. Reliability of the regressors was tested at values considered acceptable ($\rho_{xx} = .70$; Nunally, 1970, 1978), high ($\rho_{xx} = .90$), and perfect ($\rho_{xx} = 1.00$). For simplicity, the same level of regressor intercorrelation and regressor reliability was applied to all regressors in a given condition. For regressor reliability at the individual level, reliability of the regressor scores was controlled at the individual level, since most analysts assess reliability at this

level, and there is considerable disagreement about how to accurately assess reliability of aggregated reflective group-level variables (see Bliese, 2000; O’Brien, 1990). For group-level regressors, because there was no individual-level variability, reliability was controlled at the group level.

Effect size and regression coefficients The effect size was programmed at the individual regressor level in the context of the set of regressors (i.e., squared semipartial correlations). In addition to models with no effect ($f^2 = .00$), we chose to model a “medium” effect size, to ensure a valid comparison to the results of Croon and van Veldhoven (2007). Effects were modeled to corresponded to Cohen’s (1988) medium ($f^2 = .15$) effect size. For the non-null models we simulated, the regression coefficients ranged from $.10$ to $.29$ for the individual-level predictors, and from $.10$ to $.32$ for the group-level predictors. For the null models, of course, all regression coefficients were equal to zero.

ICC of the predictor variables The ICC of the predictor variables (i.e., the amount of variance located between groups) was set at $.10$ and $.20$, using the values in Croon and van Veldhoven (2007). Most work suggests that intraclass correlations in education and organizational research are usually lower than 0.30 (Bliese, 2000; Hedges & Hedberg, 2007; James, 1982). Some authors have provided guidelines for interpreting the magnitude of intraclass correlations with small, medium, and large values reported as $.05$, $.10$, and $.15$ (see Hox, 2002). As such, our selected ICC values would be considered medium and large, similar to what one might encounter in educational or organizational research.

Number of groups (N_GROUPS) We varied the number of groups on the two levels used in the Croon simulation ($N_GROUPS = 50$ and 100). To extend these values completely to what one might find in educational or organizational research we added a condition with 25 groups ($N_GROUPS = 25$).

Group size ($N1_MIN$) The number of observations in each group at the individual level was varied on four levels, based on the conditions used in Croon and van Veldhoven (2007). The first two levels kept group size fixed at either $n_j = 10$ and $n_j = 40$. In the third and fourth levels, the group sizes were varied by randomly selecting groups with small samples ranging from 5 to 15 and large samples ranging from 20 to 60, modeling unequal group sizes. A group size of 5 is normal in small group research (see Kenny, Mannetti, Pierro, Livi, & Kashy, 2002) and group sizes of 30 are typical in educational research. In multilevel research, variability in group sizes often leads to heteroscedasticity. Calculating heteroscedastic-consistent (or robust) standard errors using White correction method is often used to address this issue (see Croon & van Veldhoven, 2007; White, 1980).

The ten factors were completely crossed in the Monte Carlo study design yielding 23,328 conditions. All samples were generated from multivariate normal populations.

The research was conducted using SAS/IML version 9.1 (SAS Institute, 2004). The SAS macro provided by Hayes and Cai (2008) was used in the simulation to compute the HC3 covariance matrices for White's adjustment. Conditions for the study were run under both Windows and UNIX platforms. Normally distributed random variables were generated using the RANNOR random number generator in SAS. A different seed value for the random number generator was used in each execution of the program. The program code was verified by hand-checking results from benchmark datasets.

For each condition investigated in this study, 10,000 samples were generated. Using a large number of sample estimates allows for adequate precision in the investigation of the sampling behavior of point and interval estimates of the regression coefficients, as well as the Type I error rates and statistical power for hypothesis tests. For example, 10,000 samples provide a maximum 95% confidence interval width around an observed proportion that is $\pm .0098$ (Robey & Barcikowski, 1992).

The outcomes of interest in this simulation study included the statistical bias, standard error, and the root mean squared error (RMSE) of point estimates, as well as the Type I error control and statistical power of the hypothesis test for each coefficient. In addition, the proportions of samples that yielded inadmissible solutions for the Croon method were investigated.

Results

To guide the analysis of the Type I error control and statistical power of these tests for the regression slopes, the simulation results were analyzed using analysis of variance. The value of eta-squared (η^2) was calculated for the main effect of each research design factor in the Monte Carlo study and for their first-order interactions. Effect size η^2 is the proportion of the total variance that can be attributed to one of the factors or to an interaction between the factors. Values of η^2 greater than .0588, representing a medium effect size (Cohen, 1988), were considered large enough to merit a disaggregation of the results.

The η^2 values showed that for the individual-level predictors, the number of groups (N_GROUPS) figures prominently for both Type I error control and power for all four approaches (with η^2 ranging from .22 to .94). For the approaches based on either Croon with White's correction (CV-W) or group aggregation with White's correction (GRP-W), the number of individual-level regressors (P_X) was an important effect for Type I error control ($\eta^2 = .15$). The cross-level correlation (R_XZ) was related to the power of group aggregation

approaches both with and without White's correction (GRP) as well as to the power of the Croon (CV) approach (with η^2 ranging from .10 to .16). More importantly, we found a two-way interaction in statistical power between N_GROUPS and R_XZ for the GRP and CV approaches ($\eta^2 = .12$ and .08 for the GRP and CV approaches, respectively).

The η^2 effects were slightly lower for group-level predictors. They showed an effect on statistical power for the number of group-level regressors (P_Z) for the CV and GRP approaches ($\eta^2 = .06$ and .11, respectively), as well as the cross-level correlation (R_XZ) for the CV and GRP-W approaches ($\eta^2 = .11$ and .15, respectively). The interaction between R_XZ and N_GROUPS also produced an effect in statistical power for the CV approach ($\eta^2 = .08$). None of the remaining design factors or their interactions was important in explaining differences in Type I error and power for the various approaches.

Hypothesis tests: Type I error rates and statistical power

The distributions of the estimated Type I error rates for tests of the regression parameters of the individual-level (X) and group-level (Z) predictors are presented in Fig. 1. All four approaches provided Type I error control at or below the nominal alpha level (.05) for all conditions examined. The CV-W and GRP-W adjustments led to tests that were slightly conservative, but this effect was quite modest.

The distributions of estimated statistical power for tests of the regression parameters are presented in Fig. 2. Use of the CV and GRP methods resulted in very low power values for the tests of the regression parameters of both the individual-level predictors and the group-level predictors (power less than .10 for the majority of tests). The addition of White's adjustment to the methods (CV-W and GRP-W) resulted in a notable increase in the power of these tests (with average power near .35 for CV-W and near .65 for GRP-W).

Table 1 provides the average Type I error rates for tests of the individual-level (X) and group-level (Z) predictors. A clear effect between approaches is apparent for only number of groups (N_GROUPS) and number of individual-level predictors (P_X). The average Type I error rates are conservative for the CV, CV-W, GRP, and GRP-W analysis methods with small number of groups ($N_GROUPS = 25$). However, as the number of groups increases, error rates become closer to the nominal alpha (.05). Interestingly, there is no difference (rounded to three decimal places) in the average error rates of the X and Z predictors between any of the approaches for group size (NI_MIN) for both the individual- and group-level predictors, and for ICC for the individual-level predictors (see Table 1, panels A and B). As expected, approaches utilizing White's correction result in more conservative Type I error estimates. The number of individual-level predictors (P_X) produces error rates that become slightly more conservative as the number

Distributions of Type I Error Rates for Tests of Individual Level (X) Predictors

Distributions of Type I Error Rates for Tests of Group Level (Z) Predictors

Fig. 1 Distributions of Type I error rate estimates for the individual-level (X) and group-level (Z) predictors by analysis method

of predictors increases. As before, the CV and GRP analysis methods tend to yield less conservative Type I error rates that are closer to the nominal alpha level than do the analysis methods that utilize White's correction. Overall, the error rates for group size (NI_MIN) and ICC are slightly conservative, with the analysis methods utilizing White's correction yielding error estimates that are noticeably more stringent.

Table 2 provides average statistical power estimates for the design factors that were identified as being important. For tests of the individual-level predictors (X), the statistical power for all identified design factors is less than optimal. Over all of the identified design factors, the statistical power for the CV and GRP analysis approaches is low, whereas the approaches that utilize White's correction result in improved power. As expected, the statistical power increases as the number of groups (N_GROUPS) and the group sizes (NI_MIN) increase, with no notable differences between equal and unequal group sizes.

The CV-W and GRP-W approaches yield higher levels of statistical power than do the CV and GRP analysis methods, with the CV-W approach resulting in slightly better power. On average the power estimates for CV-W are approximately 20% higher than those for GRP-W.

Similar patterns result from the investigation of group-level predictors (Z). As the number of groups (N_GROUPS) increases, statistical power also increases. CV-W and GRP-W result in considerably improved power relative to the CV and GRP approaches, with GRP-W yielding better power levels than the CV-W analysis approach. On average, the power estimates for GRP-W are approximately 50% higher than those with CV-W. We see a more complex pattern of results for group size (NI_MIN). For the CV and GRP approaches, as group sizes increase, statistical power decreases. For the approaches that utilize White's correction, different patterns emerge. For CV-W analysis method, as group sizes increase,

Distributions of Estimated Power for Tests of Individual Level (X) Predictors

Distributions of Estimated Power for Tests of Group Level (Z) Predictors

Fig. 2 Distributions of statistical power estimates for the individual-level (X) and group-level (Z) predictors by analysis method

statistical power also increases. For the GRP-W analysis method, group size has no effect on statistical power. Across all approaches, as the numbers of individual-level (P_X) and group-level (P_Z) predictors increase, statistical power decreases.

The analysis of effects using η^2 indicated a two-way interaction in statistical power between the number of groups (N_{GROUPS}) and cross-level correlation (R_{XZ}) for the GRP and CV approaches. Figure 3 provides the average power estimates for these interactions. For the tests of both the individual-level (X) and group-level (Z) predictors, it is apparent that as the number of groups increases (from $N = 25$ to $N = 100$), statistical power improves. However, the effects of cross-level correlation on statistical power are different between the individual-level (X) and group-level (Z) predictors. For individual-level predictors (X), statistical power decreases as the cross-level correlations increase from $r_{XZ} = .0$ to $r_{XZ} = .50$.

The analysis methods utilizing White's correction result in the greatest statistical power, with CV-W yielding slightly better power performance. For tests of the group-level predictors (Z), as the cross-level correlation increases, statistical power improves. Analysis methods that utilize White's correction yield higher levels of statistical power, with the GRP-W method resulting in the highest levels. We find differing patterns of statistical power between the CV and GRP analysis methods as cross-level correlations increase, which are strengthened with White's correction. The GRP method results in increased power as cross-level correlations increase. White's correction overall yields much better power results. In contrast, for the CV analysis method, as cross-level correlations increase, statistical power remains the same or decreases as the number of groups increases. For medium to large numbers of groups, when the cross-level correlation is zero, CV-W and GRP-W yield the same amount of power. As the cross-level correlation

Table 1 Marginal mean Type I error rate estimates for tests of the individual-level (X) and group-level (Z) predictors

	CV	CV-W	GRP	GRP-W
A. Individual-level (X) predictors				
<i>Number of groups (N_GROUPS)</i>				
25	.034	.024	.034	.024
50	.043	.036	.043	.036
100	.047	.043	.047	.043
<i>Group size ($N1_MIN$)</i>				
5–15	.041	.034	.041	.034
10	.041	.034	.041	.034
20–60	.041	.034	.041	.034
40	.041	.034	.041	.034
<i>Number of individual-level predictors (P_X)</i>				
3	.042	.039	.042	.039
5	.041	.034	.041	.034
7	.041	.030	.041	.030
<i>Number of group-level predictors (P_Z)</i>				
2	.042	.037	.042	.036
4	.041	.032	.041	.032
<i>Intraclass correlation (ICC)</i>				
.1	.041	.034	.041	.034
.2	.041	.034	.041	.034
B. Group-level (Z) predictors				
<i>Number of groups (N_GROUPS)</i>				
25	.034	.024	.034	.024
50	.043	.036	.043	.036
100	.047	.043	.047	.043
<i>Group size ($N1_MIN$)</i>				
5–15	.041	.034	.041	.034
10	.041	.034	.041	.034
20–60	.041	.034	.041	.034
40	.041	.034	.041	.034
<i>Number of individual-level predictors (P_X)</i>				
3	.042	.039	.042	.039
5	.041	.034	.041	.034
7	.041	.030	.041	.030
<i>Number of group-level predictors (P_Z)</i>				
2	.042	.037	.042	.037
4	.041	.032	.041	.032

Estimates are based on 10,000 samples of each condition. Nominal $\alpha = .05$. CV = Croon & van Veldhoven, CV-W = Croon & van Veldhoven with White's adjustment; GRP = analysis conducted using group means; GRP-W = analysis conducted using group means with White's adjustment

increases, GRP-W provides power levels that are far superior to those given by the CV-W analysis method. For larger numbers of groups ($N = 50$ or 100) at the highest amount of cross-level correlations ($R_{XZ} = .50$), GRP-W yields the best statistical power, followed by GRP and CV-W. The CV analysis is the least sensitive approach when cross-level correlation is highest.

Bias of estimates

The η^2 analysis for the statistical bias estimates indicated that none of the design factors had an important effect on bias at the individual level across all of the analysis methods. For the group-level variables (Z), the only design factor that had an

important effect on bias was cross-level correlation (R_{XZ}). For the GRP and GRP-W analyses, η^2 for this effect was substantial ($\eta^2 = .15$ for both approaches). Table 3 shows the mean bias estimates for different values of cross-level correlation. The average bias for CV and CV-W was minimal, with negligible increases as cross-level correlation increased. Mean bias due to increasing cross-level correlations was more noticeable for the GRP and GRP-W analysis approaches.

Table 2 Marginal mean statistical power estimates for tests of the individual-level (X) and group-level (Z) predictors

	CV	CV-W	GRP	GRP-W
A. Individual-level (X) predictors				
<i>Number of groups (N_GROUPS)</i>				
25	.011	.105	.008	.088
50	.048	.309	.039	.264
100	.159	.589	.138	.525
<i>Group size ($N1_MIN$)</i>				
5–15	.065	.261	.048	.207
10	.066	.266	.049	.217
20–60	.079	.400	.074	.365
40	.079	.405	.075	.374
<i>Number of individual-level predictors (P_X)</i>				
3	.130	.363	.125	.367
5	.060	.365	.041	.310
7	.013	.254	.005	.168
<i>Number of group-level predictors (P_Z)</i>				
2	.104	.367	.089	.320
4	.039	.299	.033	.261
B. Group-level (Z) predictors				
<i>Number of groups (N_GROUPS)</i>				
25	.019	.159	.032	.259
50	.080	.395	.184	.634
100	.237	.675	.511	.899
<i>Group size ($N1_MIN$)</i>				
5–15	.122	.377	.294	.595
10	.115	.370	.287	.595
20–60	.110	.454	.196	.596
40	.098	.432	.188	.594
<i>Number of individual-level predictors (P_X)</i>				
3	.166	.482	.297	.646
5	.096	.389	.219	.581
7	.065	.343	.201	.551
<i>Number of group-level predictors (P_Z)</i>				
2	.152	.417	.335	.628
4	.070	.399	.146	.561

Estimates are based on 10,000 samples of each condition. Nominal $\alpha = .05$. CV = Croon & van Veldhoven, CV-W = Croon & van Veldhoven with White's adjustment; GRP = analysis conducted using group means; GRP-W = analysis conducted using group means with White's adjustment

Fig. 3 Estimated power for tests of the individual-level (X) and group-level (Z) variables by analysis method, number of groups (N_GROUPS), and cross-level correlation (R_XZ)

Although the absolute magnitude of the bias in the parameter estimates was relatively small, across all levels of R_XZ the average bias of the estimates from the GRP and GRP-W analyses (0.0323) was nearly six times as large as the average bias from the CV and CV-W analyses (0.0057), indicating that the GRP analysis methods tend to overestimate the model parameters. These results are similar to those obtained by Croon and van Veldhoven (2007).

However, an examination of the standard deviations suggests that the CV and CV-W approaches produce considerable variability in the bias estimates that is not evident for the GRP and GRP-W approaches. This variability increases as cross-level correlation increases. At the highest levels of cross-level correlation, the average standard deviation in the bias estimates for the CV and CV-W reaches 2.07. In contrast, the GRP and GRP-W methods have much less variability for the same degree of correlation—around 0.11.

These differences in the variability of statistical bias across conditions suggest that the GRP and GRP-W approaches may provide lower levels of bias across many of the conditions examined in this study. To confirm this, we compared the ratio of bias in parameter estimates from the GRP analysis to that of the CV analysis for each condition. That is,

$$\text{Bias Ratio} = \frac{\text{bias}_{GRP}}{\text{bias}_{CV}}$$

Ratios larger than 1.00 indicate conditions for which the GRP method results in more bias than the CV method, and ratios smaller than 1.00 indicate conditions for which the GRP method results in less bias. This analysis indicated that for approximately two-thirds of the conditions examined in this study (66% of conditions for the individual-level coefficients, and 69% of conditions for the group-level coefficients), the

Table 3 Mean bias estimates for coefficients of cross-level correlation (R_{XZ})

Cross-Level Correlation (R_{XZ})	CV		CV-W		GRP		GRP-W	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD
.00	– 0.003	(0.70)	– 0.003	(0.70)	0.000	(0.00)	0.000	(0.00)
.30	0.010	(0.86)	0.010	(0.86)	0.030	(0.04)	0.030	(0.04)
.50	0.010	(2.07)	0.010	(2.07)	0.067	(0.11)	0.067	(0.11)

Estimates are based on 10,000 samples of each condition. CV = Croon & van Veldhoven, CV-W = Croon & van Veldhoven with White's adjustment; GRP = analysis conducted using group means; GRP-W = analysis conducted using group means with White's adjustment

GRP method produced coefficients with less bias than those from the CV method. Our investigation of the cause of these bias results suggested that the CV and CV-W approaches induce extreme multicollinearity in some conditions. This analysis is presented in the [supplemental materials](#).

Standard errors of estimates

The standard errors of the parameter estimates from the CV, CV-W, GRP, and GRP-W analyses were also investigated. These standard errors provide an index of the sampling error associated with the parameter estimates.

From the η^2 analysis, only two of the research design factors associated with the standard errors reached an effect size that was large enough to merit a disaggregation of the results. The number of groups (N_GROUPS) was associated with standard errors for the coefficients of both the individual- and group-level predictors ($\eta^2 = .44$ and $.42$ for the individual- and group-level predictors, respectively), and the population effect size (ES) was also associated with the coefficients of the predictors at both levels ($\eta^2 = .36$ and $.41$ for the individual- and group-level predictors, respectively). The marginal mean values of the standard errors for these factors are provided in Table 4. As expected, the standard errors decreased with larger numbers of groups and larger effect sizes. However, the difference in standard errors between the two CV methods and the two GRP methods is striking. For the coefficients of both the individual- and group-level predictors, the mean standard errors for the GRP and GRP-W analyses remained well below 1.00. In contrast, for the CV and CV-W analyses, the mean standard error ranged from 19.08 to 87.60 for the individual-level predictors, and from 18.74 to 67.46 for the group-level predictors, indicating that the CV approach produces much less precise model parameter estimates with fewer groups and small effect sizes.

Root mean squared error (RMSE) of estimates

In addition to the examination of bias in the point estimates, the root-mean squared error (RMSE) of the estimates was examined. This statistic provides an index of the total error

in the parameter estimates, combining both statistical bias and sampling error.

From the η^2 analysis, only two of the research design factors associated with RMSE reached an effect size that was large enough to merit a disaggregation of the results. The number of groups (N_GROUPS) was associated with RMSE for the coefficients of both the individual- and group-level predictors ($\eta^2 = .49$ and $.41$ for the individual- and group-level predictors, respectively), and cross-level correlation (R_{XZ}) was associated with RMSE for the coefficients of the group-level predictors ($\eta^2 = .08$). The marginal mean values of RMSE for these factors are provided in Table 5. As expected, the RMSE decreased with larger numbers of groups, and increased with higher levels of cross-level correlation. However, the difference in RMSEs between the two CV methods and the two GRP methods is noticeable. For the coefficients of both the individual- and group-level predictors, the mean RMSE for the GRP and GRP-W analyses remained well below 1.00. In contrast, for the CV and CV-W analyses, the mean RMSE ranged from 18.70 to 83.22 for the individual-level predictors, and from 19.19 to 67.62 for the group-level predictors. These results indicate that the total error (combining accuracy and precision) of the CV and CV-W analyses was notably larger than that from GRP and GRP-W.

Predicting inadmissible solutions

Croon and Van Veldhoven (2007) noted that their approach occasionally yielded inadmissible solutions in their simulations. Such inadmissible solutions are typical with iterative procedures, such as mixed models and structural equation modeling. We found similar results: Inadmissible solutions, or model nonconvergence, were obtained in 1,572 conditions (approximately 7% of the conditions simulated). To analyze the probabilities of obtaining inadmissible solutions, we conducted a logistic regression analysis with the admissibility of the solution as the binary outcome and the simulation design factors as regressors. The results of the logistic regression analysis are provided in Table 6. As is evident in this table, the probability of obtaining an inadmissible solution increased

Table 4 Marginal mean standard error estimates for coefficients of individual-level (*X*) and group-level (*Z*) predictors

	CV	CV-W	GRP	GRP-W
<i>A. Individual-level (X) predictors</i>				
<i>Number of groups (N_GROUPS)</i>				
25	87.60	87.60	0.20	0.20
50	42.13	42.13	0.12	0.12
100	19.08	19.08	0.08	0.08
<i>Effect size (ES)</i>				
0.00	58.29	58.29	0.18	0.18
0.15	40.27	40.27	0.09	0.09
<i>B. Group-level (Z) predictors</i>				
<i>Number of groups (N_GROUPS)</i>				
25	67.46	67.46	0.23	0.23
50	30.89	30.89	0.14	0.14
100	18.74	18.74	0.09	0.09
<i>Effect size (ES)</i>				
0.00	47.93	47.93	0.22	0.22
0.15	32.13	32.13	0.11	0.11

Estimates are based on 10,000 samples of each condition. Nominal $\alpha = .05$. CV = Croon & van Veldhoven, CV-W = Croon & van Veldhoven with White's adjustment; GRP = analysis conducted using group means; GRP-W = analysis conducted using group means with White's adjustment

with larger effect size (ES), higher levels of ICC, higher levels of correlation between the group-level predictors (R_Z), and higher cross-level correlations (R_{XZ}). In addition, a substantial interaction between the group-level predictor correlation and cross-level correlation was obtained ($R_Z * R_{XZ}$). To facilitate interpretation of this interaction, the interaction component of the logistic model is graphed in Fig. 4. Evident in this figure is that the increase in the probability of program failure (inadmissible solutions) occurs when the cross-level correlation is high ($R_{XZ} = .50$) and the correlations between the predictors at the group level are low ($R_Z = .20$ or $.40$).

The sample percentages of inadmissible solutions are presented in Table 7. We see an increase in inadmissible solutions when the cross-level correlation (R_{XZ}) is larger than the correlation between the group-level variables (R_Z), and the percentages increase with higher levels of ICC.

Full information maximum likelihood estimation

Lüdtke et al. (2008) described a full-information maximum likelihood estimation method (FIML) for analyses that use individual averages as the group-level predictors in multilevel models (also known as *modeling contextual effects*). Although they demonstrated this approach on multilevel data with a dependent variable measured at the individual rather than the group level, a comparative investigation that is specific to data

instances in which the dependent variable is measured at the group level would provide an important extension of the work of both Croon and van Veldhoven (2007) and Lüdtke et al. (2008). As such, we conducted an additional simulation with a partial replication of the full simulation design to compare this FIML method to the Croon and group-mean analysis methods. The FIML method was found to have poor control of Type I error probabilities in the majority of conditions examined, and severe problems with nonconvergence. Details about this simulation study are provided in the [supplemental materials](#).

Discussion

This comparison of analytic strategies for group-level outcomes suggests that little is gained with the Croon and van Veldhoven (2007) approach relative to an OLS analysis of group means in conjunction with White's adjustment for heteroscedasticity. Type I error rates were the same for the group-level analysis (GRP) and the method recommended by Croon and van Veldhoven (CV). Differences between the approaches were more evident with statistical power. The GRP analysis showed substantially lower power than the CV-W analysis. Power for the group means analysis was improved by utilizing White's adjustment for heteroscedasticity, although results compared to the CV-W approach were not consistently superior. As compared to an analysis of group means in conjunction with White's adjustment for heteroscedasticity (GRP-W), CV-W evidenced slightly greater power for testing the individual-level (*X*) predictors and substantially lower power for testing the coefficients of the group-level (*Z*) predictors. We also found a significant interaction in statistical power between the number of groups (*N_GROUPS*) and cross-level correlation (R_{XZ}) that differs for individual- and group-level predictors. For individual-level (*X*) predictors, as *N_GROUPS* increase, statistical power improves for both approaches when White's correction is employed. As the cross-level correlation increases, however, statistical power decreases. For group-level (*Z*) predictors using the GRP approaches, as *N_GROUPS* and R_{XZ} increased, statistical power improves. For group-level predictors using the CV approaches, statistical power decreases as *N_GROUPS* and R_{XZ} increase. For both approaches, the power magnitude and impact of the interaction is more prominent when White's correction is utilized.

Consistent with the results of Croon and van Veldhoven (2007), the GRP analyses yielded parameter estimates that were more biased, on average, than those obtained with the CV analyses but the absolute magnitude of these biases were relatively small. However, the CV strategy of computing adjusted means for the individual-level predictors increased the level of multicollinearity in the samples with a concomitant

Table 5 Marginal mean RMSE estimates for tests of individual-level (X) and group-level (Z) predictors

	CV	CV-W	GRP	GRP-W
<i>A. Individual-level (X) predictors</i>				
<i>Number of groups (N_GROUPS)</i>				
25	83.22	83.22	0.21	0.21
50	41.82	41.82	0.14	0.14
100	18.70	18.70	0.10	0.10
<i>Group-level (Z) predictors</i>				
<i>Number of groups (N_GROUPS)</i>				
25	63.67	63.67	0.26	0.26
50	33.51	33.51	0.17	0.17
100	19.19	19.19	0.12	0.12
<i>Cross-level correlation (R_XZ)</i>				
0.00	30.44	30.44	0.17	0.17
0.30	67.62	67.62	0.22	0.22
0.50				

Estimates are based on 10,000 samples of each condition. Nominal $\alpha = .05$. CV = Croon & van Veldhoven, CV-W = Croon & van Veldhoven with White's adjustment; GRP = analysis conducted using group means; GRP-W = analysis conducted using group means with White's adjustment

increase in statistical bias for these conditions. In addition, the CV strategy produced inadmissible solutions for approximately 7% of the samples in the simulation study (a result also noted by Croon & van Veldhoven, 2007). These inadmissible solutions were most frequently encountered in samples with larger effect sizes, higher levels of intraclass correlation, and cross-level correlations that are higher than the correlations between group-level predictors.

Finally, from a practical perspective on implementation, the CV strategy is not available in the major statistical packages used by applied researchers. Croon and van Veldhoven (2007) provided an S-Plus script for computation and the present authors programmed the computations in SAS/IML and MPLUS (Muthén & Muthén, 2007), making the strategy available for researchers who use these languages (see the [supplemental materials](#) and https://osf.io/4rgvh/?view_only=133543b6151a4ccbbde895839ceef378). In contrast, the GRP strategy is easily implemented with any package that performs OLS regression.

Although our general recommendation for the researcher is to rely on a GRP analysis combined with White's correction, we acknowledge that the differences in power performance between the two approaches may temper our endorsement. If the researcher is primarily interested in group-level predictors (Z), then using the GRP aggregation approach in conjunction with White's correction will maximize statistical power for these predictors. If the focus is on individual-level (X) predictors, then our results suggest that using the CV approach followed by White's correction yields somewhat better power rates. Overall, however, we find that statistical power for predictors at the individual level (X) only approaches acceptable levels (power = .80) with little to no cross-level correlation and at least 100 groups, combined with White's correction. For predictors at the group level (Z), using the GRP approach in conjunction with White's correction results in the greatest statistical power. In combination with the GRP-W approach, numbers of groups as small as 50 yield adequate statistical power when associated with moderate cross-level correlations.

Table 6 Logistic regression predicting program failure

Predictor	β	$SE(\beta)$	Wald χ^2	Odds Ratio	95% CI (Odds Ratio)	
					Lower	Upper
Intercept	− 12.786	0.44	826.086			
ES	13.965	0.51	750.721	>999.999	>999.999	>999.999
ICC	3.879	0.59	42.833	48.395	15.144	154.651
NI_MIN	− 0.001	0.00	0.058	0.999	0.995	1.004
N_GROUPS	0.003	0.00	7.586	1.003	1.001	1.004
P_X	0.358	0.02	345.244	1.430	1.377	1.485
P_Z	0.002	0.03	0.003	1.002	0.946	1.061
R_X	− 1.136	0.18	39.277	0.321	0.225	0.458
R_XX	0.594	0.24	6.229	1.811	1.136	2.886
R_XZ	17.495	0.70	620.194	>999.999	>999.999	>999.999
R_Z	8.166	0.61	177.719	>999.999	>999.999	>999.999
$R_Z * R_XZ$	− 26.900	1.39	373.451	0.001	0.001	0.001

$n = 23,328$; ES = effect size; ICC = intraclass correlation; NI_MIN = group size; N_GROUPS = number of groups; individual (X) predictor factors: P_X = number of X s; R_X = correlation between X s; R_XX = reliability of X s; group (Z) predictor factors: P_Z = number of Z s; R_Z = correlation between Z s; R_XZ = correlation between X and Z s

Program Failure by Cross- and Group-Level (R_Z) Correlations

Fig. 4 Probabilities of program failure by cross-level correlations (R_{XZ}) and correlation between the group-level predictors (R_Z)

Although some may be tempted to adopt a hybrid approach in order to exploit the power advantages of each analysis strategy (i.e., using CV-W for testing individual-level effects and GRP-W for testing group-level effects), we do not advocate such a course. Pursuing such parallel approaches increases the complexity of the data analysis, as well as the requisite explanations needed to describe and justify such analyses. In addition, the person-as-variables approach and the simple-group-means approach represent disparate philosophical views of multilevel data and the processes underlying them. Finally, the CV approach provides adjustments to both individual-level and group-level regressors (see Eq. 1.4). Allowing such adjustments to affect one set of tests while ignoring them for another set of tests represents a level of statistical *ad hocery*—that is awkward at best.

Table 7 Percentages of samples with inadmissible solutions by effect size, ICC, group-level correlation, and cross-level correlation

R_Z	R_{XZ}	Effect Size = 0		Effect Size = .15	
		ICC = .1	ICC = .2	ICC = .1	ICC = .2
.2	.0	0%	0%	0%	0%
	.3	0%	4%	7%	9%
	.5	2%	6%	44%	51%
.4	.0	1%	0%	0%	0%
	.3	2%	4%	6%	6%
	.5	0%	6%	22%	24%
.6	.0	0%	1%	6%	6%
	.3	2%	4%	6%	6%
	.5	1%	5%	7%	8%

ICC = intraclass correlation; R_Z = correlation between Z s; R_{XZ} = correlation between X and Z s

In general, the number of regressors at the individual or group levels was not an important consideration in statistical power or error control. For both the individual and group levels, a greater number of regressors was associated with slight reductions in statistical power and increases in Type I error rates (see Kromrey & Foster-Johnson, 1999b). Contrary to other studies (see Kromrey et al., 2007), the reliability of the regressors did not have an effect on Type I error rates, statistical power, or bias estimates in our analysis. It is possible that the lack of effects may be due to the relatively high levels of regressor reliability used in this study or the method used to model regressor reliability may not adequately capture the complexity of measurement error in a multilevel context (see Raykov & du Toit, 2005; Raykov & Marcoulides, 2006; Raykov & Penev, 2009). Future work should expand the range of regressor reliability values and explore a more sophisticated method of generating multilevel regressor reliabilities.

This study is not without limitations. The range of effect sizes should be broadened to include small and large effect sizes, and a full spread of ICC values should be explored. In addition, the group sizes modeled in this study are somewhat contrived and could be improved by programming the naturally occurring variability often encountered in a “real-world” environment. This investigation was also limited to linear regression equations—an examination of the comparative performance of the CV and GRP approaches with non-linear models would be informative. Additionally, all of the variables in our models were based on normal distributions; knowing the relative performance of CV and GRP approaches on regressors with non-normal distributions would contribute to our understanding of these methods.

The approach recommended by Croon and van Veldhoven (2007), Lüdtke et al. (2008), and Bennink, Croon, and Vermunt (2013) incorporate fundamental components of the

persons-in-variables approach, in which the correlations within and between levels are explicitly acknowledged and accounted for. This philosophy should not be completely disregarded—a standard aggregation analysis approach that ignores these correlations does not accurately capture the complexities of a multilevel data structure. If the persons-in-variables approach resonates with the researcher, then we may offer a cautious recommendation for the use of CV, accompanied with a warning to attend carefully to the potentially problematic data structures identified in this study.

In selecting an analysis strategy, the recommendations of Wilkinson and the Task Force on Statistical Inference (1999) should be considered:

The enormous variety of modern quantitative methods leaves researchers with the nontrivial task of matching analysis and design to the research question. Although complex designs and state-of-the-art methods are sometimes necessary to address research questions effectively, simpler classical approaches often can provide elegant and sufficient answers to important questions. Do not choose an analytic method to impress your readers or to deflect criticism. If the assumptions and strength of a simpler method are reasonable for your data and research problem, use it. Occam's razor applies to methods as well as to theories. (p. 598)

Author note The authors express appreciation for the helpful feedback provided by the editor and anonymous reviewers. The manuscript was much improved by following their suggestions. The authors also thank Linda Muthén for her invaluable assistance with the MPLUS syntax. Finally, the authors acknowledge the support of the Dartmouth Research Computing Center for providing access to high-speed computing resources.

References

- Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics*, 6, 267–285.
- Bauer, D. J. (2003). Estimating multilevel linear models as structural models. *Journal of Educational and Behavioral Statistics*, 28, 135–167.
- Bell-Ellison, B. A., Ferron, J. M., & Kromrey, J. D. (2008). Cluster size in multilevel models: The impact of small level-1 units on point and interval estimates in two level models. In Proceedings of the American Statistical Association, Social Statistics Section [CD-ROM], Alexandria: American Statistical Association.
- Bennink, M., Croon, M. A., & Vermunt, J. K. (2013). Micro-macro multilevel analysis for discrete data: A latent variable approach and an application on personal network data. *Sociological Methods and Research*, 42, 431–457.
- Bliese, P., Chan, D., & Ployhart, R. (2007). Multilevel methods: Future directions in measurement, longitudinal analyses, and non-normal outcomes. *Organizational Research Methods*, 10, 551–563.
- Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco: Jossey-Bass.
- Cai, L., & Hayes, A. F. (2008). A new test of linear hypotheses in OLS regression under heteroscedasticity of unknown form. *Journal of Educational and Behavioral Statistics*, 33, 21–40.
- Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, 83, 234–246.
- Chan, D. (2005). Multilevel research. In F. T. L. Leong & J. T. Austin (Eds.), *The psychology research handbook* (2nd, pp. 401–418). Thousand Oaks: Sage.
- Clark, W. A. V., & Avery, K. L. (1976). The effects of data aggregation in statistical analysis. *Geographical Analysis*, 8, 428–438.
- Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health*, 62, 752–758.
- Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research using cluster analysis to create synthetic neighborhoods. *Sociological Methods and Research*, 35, 311–351.
- Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics*, 10, 637–666.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd). Hillsdale: Erlbaum.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale: Erlbaum.
- Cribari-Neto, F., Ferrari, S. L. P., & Cordeiro, G. M. (2000). Improved heteroscedasticity-consistent covariance matrix estimators. *Biometrika*, 87, 907–918.
- Cribari-Neto, F., Ferrari, S. L. P., & Oliveira, W. A. S. C. (2005). Numerical evaluation of tests based on different heteroskedasticity-consistent covariance matrix estimators. *Journal of Statistical Computation & Simulation*, 75, 611–628.
- Cribari-Neto, F., & Zarkos, S. G. (2001). Heteroscedasticity-consistent covariance matrix estimation: White's estimator and the bootstrap. *Journal of Statistical Computation and Simulation*, 68, 391–411.
- Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, 12, 45–57.
- Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, 38, 529–569.
- Davies, R. B., & Hutton, B. (1975). The effect of errors in the independent variables in linear regression. *Biometrika*, 62, 383–391.
- Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., . . . Lee, R. L. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, 79, 69–102.
- Fisher, E. S., Bynum, J. P., & Skinner, J. S. (2009). Slowing the growth of health care costs—Lessons from regional variation. *New England Journal of Medicine*, 360, 849–52.
- Hayes, A. F., & Cai, L. (2008). Using heteroscedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, 39, 709–722.
- Hedeker, D., Gibbons, R. D., & Flay, B. R. (1994). Random-effects regression models for clustered data with an example from smoking prevention research. *Journal of Consulting and Clinical Psychology*, 62, 757–765.
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60–87.

- Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management*, 23, 723–744.
- Hopkins, K. D. (1982). The unit of analysis: Group means versus individual observations. *American Educational Research Journal*, 19, 5–18.
- Hox, J. (2002). *Multilevel analysis: Techniques and application*. Mahwah, Lawrence Erlbaum.
- Hutchison, D. (2003). Bootstrapping the effect of measurement errors on apparent aggregated group-level effects. In S. P. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 209–228). Mahwah: Erlbaum.
- Huynh, C. L. (2006). Estimation and diagnostics for errors-in-variables mixed effects models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin*, 117, 348–357.
- James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology*, 67, 219–229.
- Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology*, 83, 126–137. doi:<https://doi.org/10.1037/0022-3514.83.1.126>
- Kozlowski, S. W., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. Kowloski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco: Jossey-Bass.
- Kromrey, J., & Foster-Johnson, L. (1998). Mean centering in moderated multiple regression: Much ado about nothing. *Educational and Psychological Measurement*, 58, 42–67.
- Kromrey, J., & Foster-Johnson, L. (1999a). Statistically differentiating between interaction and nonlinearity in multiple regression analysis: A Monte Carlo investigation of a recommended strategy. *Education and Psychological Measurement*, 59, 392–413.
- Kromrey, J., & Foster-Johnson, L. (1999b). Testing weights in multiple regression analysis: An empirical comparison of protection and control strategies. Paper presented at the Joint Meetings of the American Statistical Association, Baltimore.
- Kromrey, J., Yi, Q., & Foster-Johnson, L. (1997). Statistical bias resulting from the use of variable selection algorithms in discriminant function analysis: What do stepwise-build models represent? Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.
- Kromrey, J. D., Coraggio, J. T., Phan, H., Romano, J. T., Hess, M. R., Lee, R. S., ... Luther, S. L. (2007). Fallible regressors in multilevel models: Impact of measurement error on parameter estimates and hypothesis tests. Paper presented at the American Educational Association's Annual Meeting, Chicago.
- Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician*, 54, 217–224.
- Longford, N. (1993). Regression analysis of multilevel data with measurement error. *British Journal of Mathematical and Statistical Psychology*, 46, 301–311.
- Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13, 203–229.
- Maas, C. J. M., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, 46, 427–440.
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1, 86–92.
- MacKinnon, J. G., & White, H. (1985). Some heteroscedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29, 305–325.
- Maxwell, S. E., Delaney, H. D., & Dill, C. A. (1984). Another look at ANCOVA versus blocking. *Psychological Bulletin*, 95, 136–147.
- Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, 10, 259–284.
- Mok, J. (1995). *Sample size requirements for 2-level designs in educational research*. Unpublished manuscript, Macquarie University, Sydney, Australia.
- Mossholder, K. W., & Bedeian, A. G. (1983). Cross-level inference and organizational research: Perspectives on interpretation and application. *Academy of Management Review*, 8, 547–558.
- Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide*. Los Angeles: Muthén & Muthén.
- Myer, A., Thoroughgood, C., & Mohammed, S. (2016). Complementary or competing climates? Examining the interactive effect of service and ethical climates on company-level financial performance. *Journal of Applied Psychology*, 101, 1178–1141.
- Newsom, J. T., & Nishishiba, M. (2002). *Nonconvergence and sample bias in hierarchical linear modeling of dyadic data*. Unpublished manuscript, Portland State University.
- Nunnally, J. C. (1970). *Introduction to psychological measurement*. New York: McGraw-Hill.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- O'Brien, R. M. (1990). Estimating the reliability of aggregate-level variables based on individual-level characteristics. *Sociological Methods and Research*, 18, 473–504.
- Pedhazuer, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale: Erlbaum.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park: Sage.
- Raykov, T., & du Toit, S. H. C. (2005). Estimation of reliability for multiple-component measuring instruments in hierarchical designs. *Structural Equation Modeling*, 12, 536–550.
- Raykov, T., & Marcoulides, G. A. (2006). On multilevel model reliability estimation from the perspective of structural equation modeling. *Structural Equation Modeling*, 13, 130–141.
- Raykov, T., & Penev, S. (2009). Estimation of maximal reliability for multiple component instruments in multilevel designs. *British Journal of Mathematical and Statistical Psychology*, 62, 129–142.
- Richter, F. G. C., & Brorsen, B. W. (2006). Aggregate versus disaggregate data in measuring school quality. *Journal of Production Analytics*, 25, 279–289.
- Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of interactions in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, 45, 283–288.
- SAS Institute Inc. (2004). *SAS, release 9.12* [Computer program]. Cary: SAS Institute, Inc.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.
- Thompson, B. (1990). Finding a correction for the sampling error in multivariate measures of relationship: A Monte Carlo study. *Educational and Psychological Measurement*, 50, 15–31.
- Thompson, B. (1991). Invariance of multivariate results: A Monte Carlo study of canonical function and structure coefficients. *Journal of Experimental Education*, 59, 367–382.
- Van Mierlo, H., Vermunt, J. K., & Rutte, C. G. (2009). Composing group-level constructs from individual-level survey data. *Organizational Research Methods*, 12, 368–392.

- White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, 48, 817–838.
- Wilkinson, L., & the Task Force on Statistical Significance. (1999). Statistical methods in psychology journals. *American Psychologist*, 54, 594–604.
- Wood, S., Van Veldhoven, M., Croon, M., & de Menezes, L. M. (2012). Enriched job design, high involvement management and organizational performance: The mediating roles of job satisfaction and well-being. *Human Relations*, 65, 419–446.
- Woodhouse, G., Yang, M., Goldstein, H., & Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society: Series A*, 159, 201–212.