Group-Level Outcomes in Multilevel Designs: An Empirical Comparison of Analysis Strategies

Jeffrey D. Kromrey[1] and Lynn Foster-Johnson[2]
[1]University of South Florida, Tampa, FL
[2]Geisel School of Medicine, Dartmouth College, Hanover, NH

Abstract

Methods for analyzing multilevel data with group-level outcome variables were compared in a simulation study. The analytical methods included OLS analyses of group means, a two-step approach suggested by Croon and van Veldhoven (2007), and a Full Information Maximum Likelihood Latent variable technique proposed by Lüdtke et al. (2008). Type I error control, power, bias, standard errors, and RMSE in parameter estimates were compared across design conditions that included number of predictor variables, level of correlation between predictors, level of intraclass correlation, predictor reliability, effect size, and sample size. Results suggested that an OLS analysis of 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. Further, this simple analysis avoided the extreme bias in parameter estimates and inadmissible solutions that were encountered with other strategies. Results were interpreted in terms of recommended analytical methods for applied researchers.

Key Words: Simulation, Type I Error, Bias, Power, Multilevel

1. 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.

A substantial body of work has been developed for analyzing multilevel data where the outcome variable is measured at the individual level (see, for example, Raudenbush & Bryk, 2002). Often referred to as the *macro-micro* data situation (cf., Snijders & Bosker, 1999), 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 (L2) 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, where Y is measured at the higher (group) level and corresponding explanatory variables are measured at the individual level (X) and at the group level (Z). Generally, there have been two approaches for analyzing micro-macro data. While 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 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.

1.1 A 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 where the relationship between the group scores on explanatory variables Z and ξ, and the outcome variable Y is:

$$y_g = \beta_0 + \beta_1 \xi_g + \beta_2 Z_g + \in_g$$

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 a reflective indicator for the unobserved group score. 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 $\in$ is assumed to be homoscedastic, or to have a constant variance for all groups.

All three parameters in the equation above are defined at the group level, but because 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 is denoted by σ^2_ξ, and the variance of the disturbance term ν_{ig} (assumed to be constant for all subjects and groups) by σ^2_ν. The variance σ^2_ξ is the between-group variance of X; the variance σ^2_ν is its within-group variance.

1.2 A Full Information Maximum Likelihood Approach

Lüdtke et al. (2008) described a full information maximum likelihood estimation method (FIML) for the analysis of contextual effects in multilevel models. Although this approach was demonstrated on multilevel data with a dependent variable measured at the individual level, a comparative investigation that is specific to data instances where the

dependent variable is measured at the group level would be an important extension of the work of both Croon and van Veldhoven (2007) and Lüdtke et al. (2008) and may also provide an alternative to the recommended analysis approaches. Understanding how the FIML approach performs under those data conditions where both the current investigation and the work of Croon and Van Veldhoven (2007) yielded unacceptable results or model non-convergence would lead to more precise recommendations of the best approach to employ with intractable data configurations. 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.

2. Purpose of the Study

The primary purpose of this 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. In a separate analysis (Study 2), we also investigated the comparative performance of the Full Information Maximum Likelihood (FIML) approach recommended by Lüdtke at al. (2008). A comparative investigation that is specific to data instances where the dependent variable is measured at the group level (L2) is an important extension of the work of both Croon and van Veldhoven (2007) and Lüdtke et al. (2008) and may provide an alternative to the typical aggregation approach currently used with data configurations such as these.

3. Method (Study 1)

The statistical performance of the Croon method (with [CV-W] and without White's adjustment [CV]), and a traditional regression analysis using group means (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 reflective indicators of the group level construct, where the scores associated with individuals in a group are interchangeable.

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. At the individual level, we included models with 3, 5, and 7 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, we included models with 2 and 4 group-level regressors.

Correlation between individual- and group-level regressors. We varied the correlation between the individual regressors by levels that would be considered low, medium, and

high inter-regressor correlations (ρ_X = .10, .30, and .50). Correlations between group level regressors were varied by values of (ρ_Z = .20, .40, and .60).

Cross-level correlations. Cross-level correlations were set to zero, moderate, and high correlations (ρ_{XZ} = 0, .30, and .50). These values allowed comparison to Croon and van Veldhoven (2007) as well as providing performance information on a scenario where cross-level correlations were high.

Reliability of Regressors. 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 (ρ_{XX} = .70; Nunally, 1970; 1978), high (ρ_{XX} = .90), and perfect (ρ_{XX} = 1.00). For simplicity, the same level of regressor intercorrelation and regressor reliability was applied to all regressors in a given condition. 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 (cf., Bliese, 2000; O'Brien, 1990).

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 semi-partial correlations). In addition to models with no effect (f^2 = 0.00), we chose to model a "medium" effect size, to ensure a valid comparison to the results of Croon & van Veldhoven (2007). Effects were modeled to corresponded to Cohen's (1988) medium (f^2 = 0.15) effect size. For the non-null models we simulated, the regression coefficients ranged from 0.10 to 0.29 for the individual level predictors and from 0.10 to 0.32 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 & 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 (cf., 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. We varied the number of groups on the two levels used in the Croon simulation (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.

Group size. The number of observations in each group at the individual level was varied on three 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 level, group sizes were varied by randomly selecting groups with small samples ranging from 5 to 15 and large samples ranging from 20 to 60. A group size of 5 is normal in small group research (cf., Kenny, Kashy, Mannetti, Pierro, & Levi, 2002) and group sizes of 30 are typical in educational research. In multilevel research, variability in group sizes often leads to heteroscdasticity. Calculating heteroscedastic-consistent (or robust) standard errors using White correction method is often used to address this issue (cf., Croon & VanVeldhoven, 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).

4. Results (Study 1)

4.1 Type I Error Control

The distributions of the estimated Type I error rates for the tests of regression parameters of the Individual level (X) and Group level (Z) predictors are presented in Figure 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.

4.2 Statistical Power

The distributions of estimated statistical power for the tests of the regression parameters are presented in Figure 2. The use of CV and GRP 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). Further, differences in power between CV-W and GRP-W were noted for the tests of Individual level and Group level predictors. The CV-W approach provided slightly

greater power for tests of Individual level predictors, but notably lower power for tests of Group level predictors.

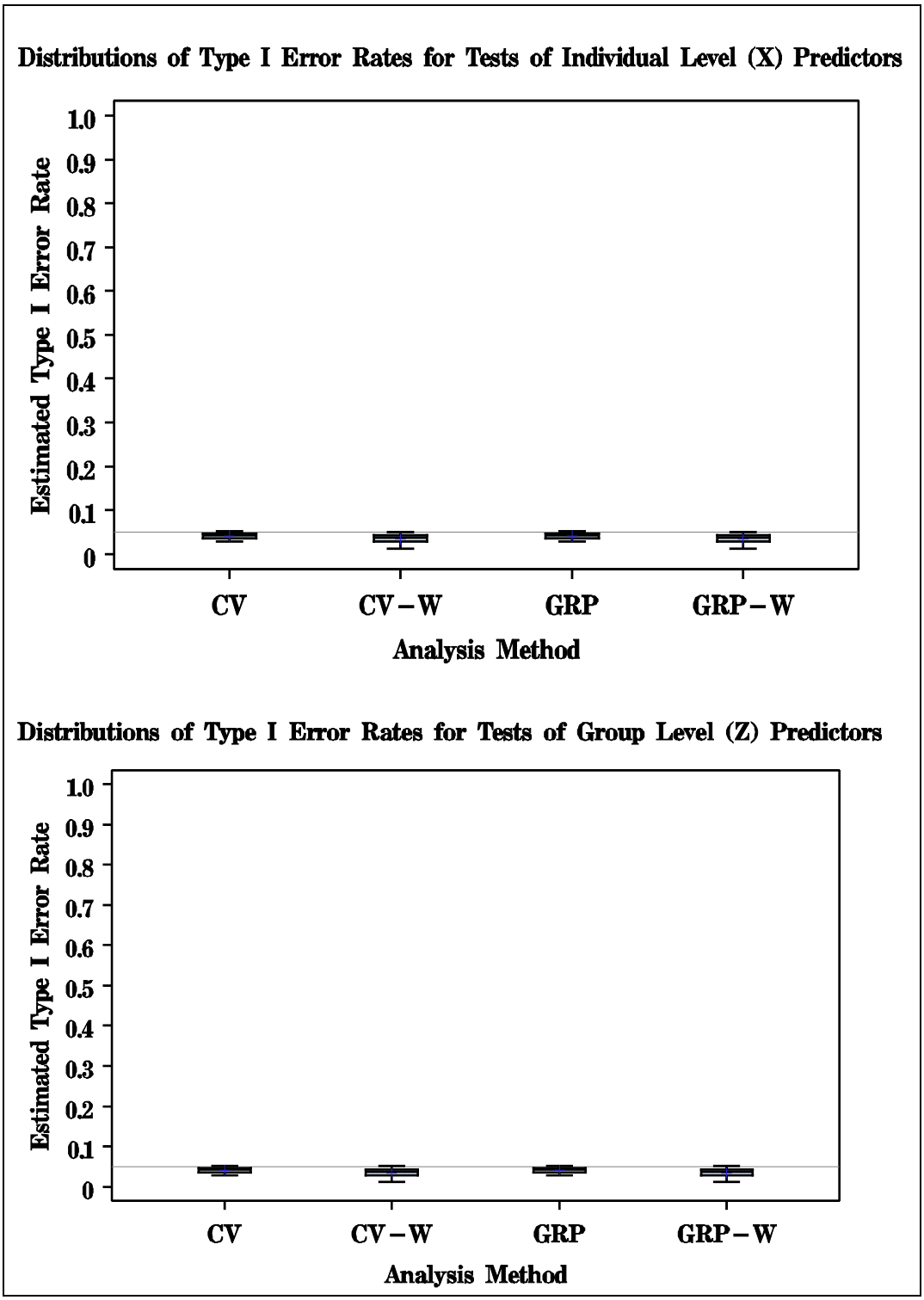

Figure 1: Distributions of Estimated Type I Error Rates Across Study Conditions (α = .05)

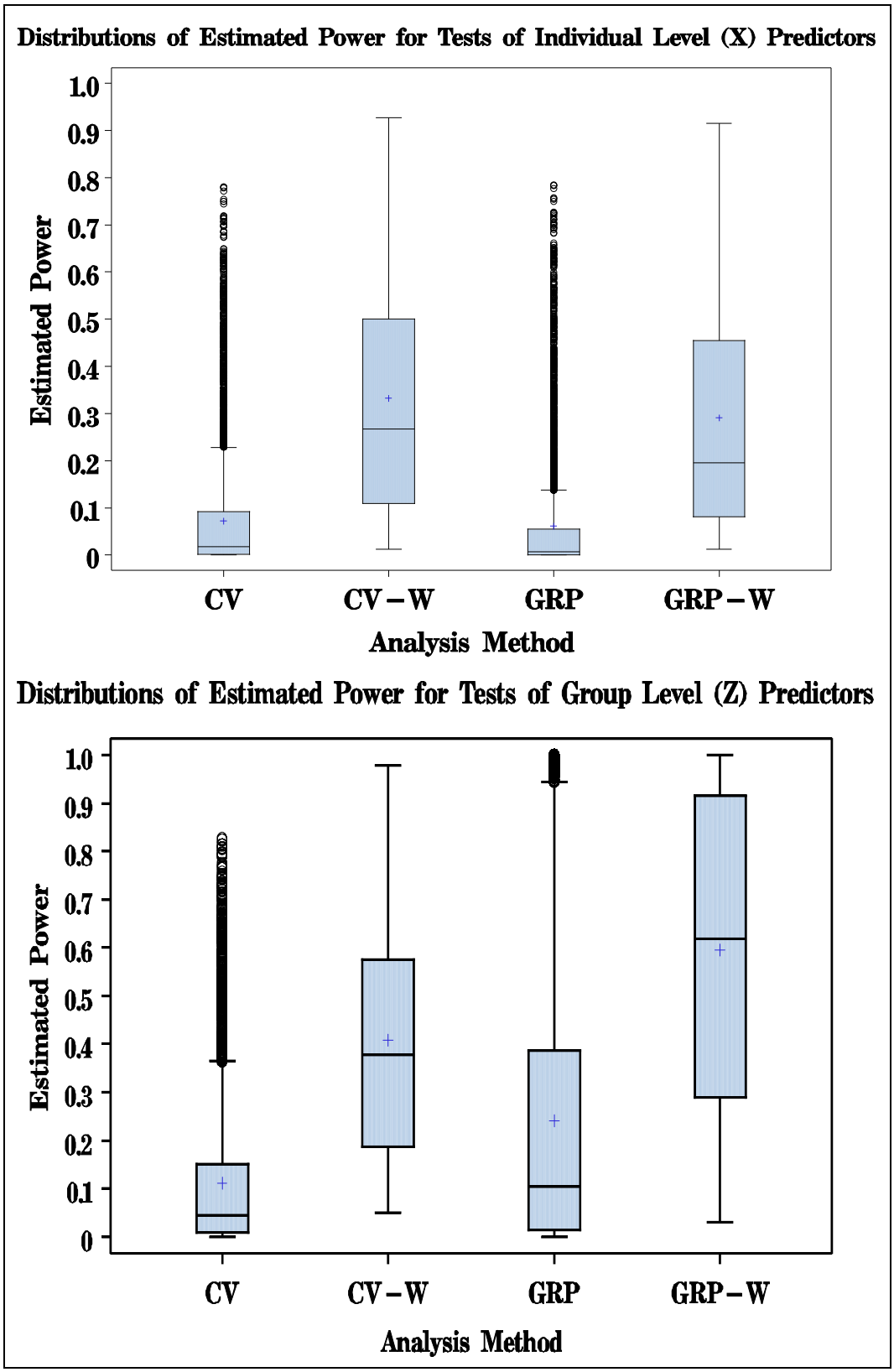

Figure 2: Distributions of Estimated Statistical Power Across Study Conditions (α = .05)

4.3 Bias in Parameter Estimates

The distributions of estimated bias in parameter estimates are provided in Figure 3. The average bias was near zero for all analyses. However, the results indicate that for the CV and CV-W approaches, there is 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 correlations, 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. Our investigation of the cause of these bias results (not reported here) suggest that the CV and CV-W approaches induce extreme multicollinearity in some conditions.

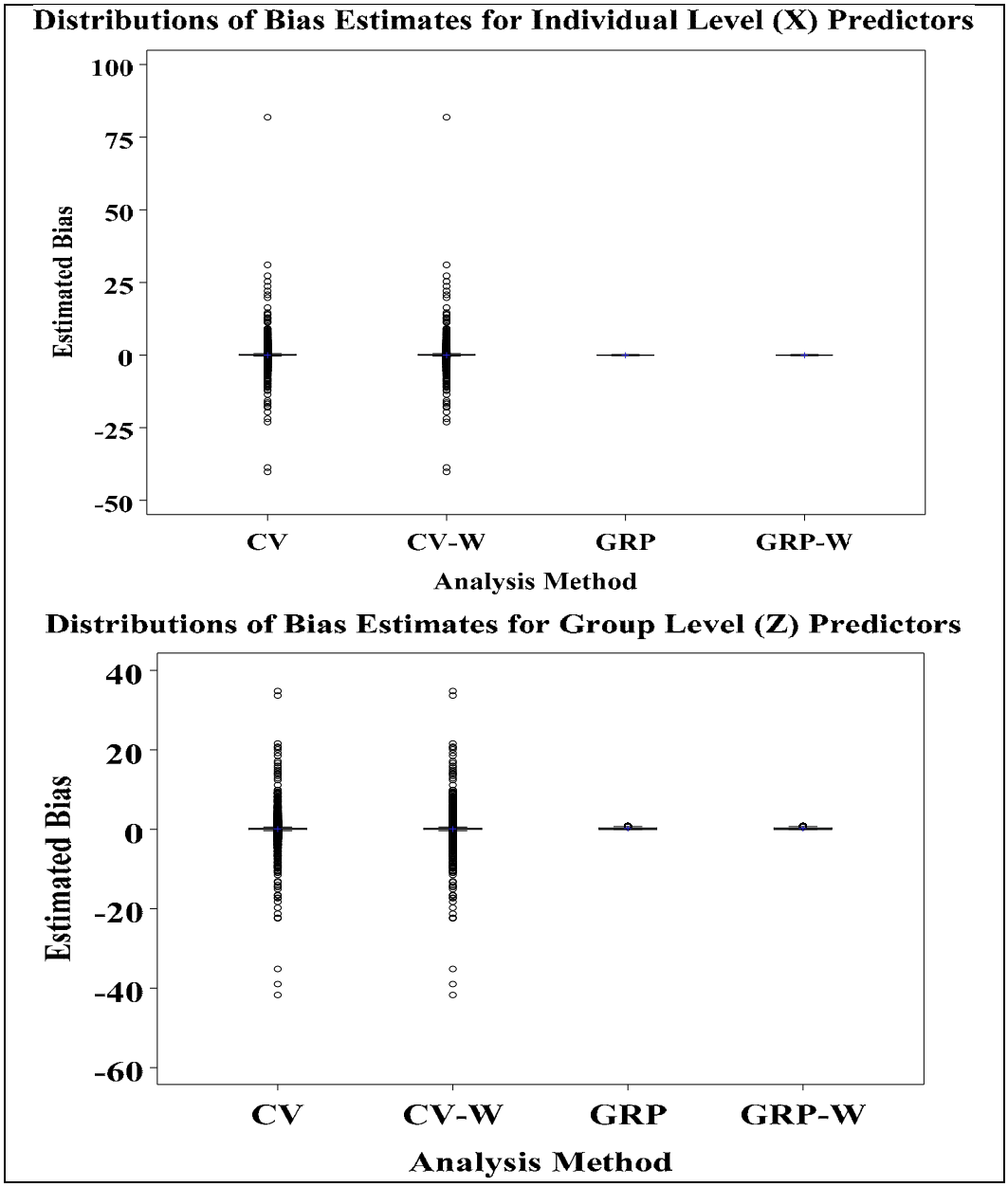

Figure 3: Distributions of Estimated Bias Across Study Conditions (α = .05)

4.4 RMSE in Parameter Estimates

The distributions of Root Mean Square Error (RMSE) in parameter estimates are provided in Figure 4. Concomitant with the variability in bias, large variability in RMSE is evident for the CV and CV-W approaches. In contrast, the use of GRP and GRP-W approaches result in consistently small RMSE estimates.

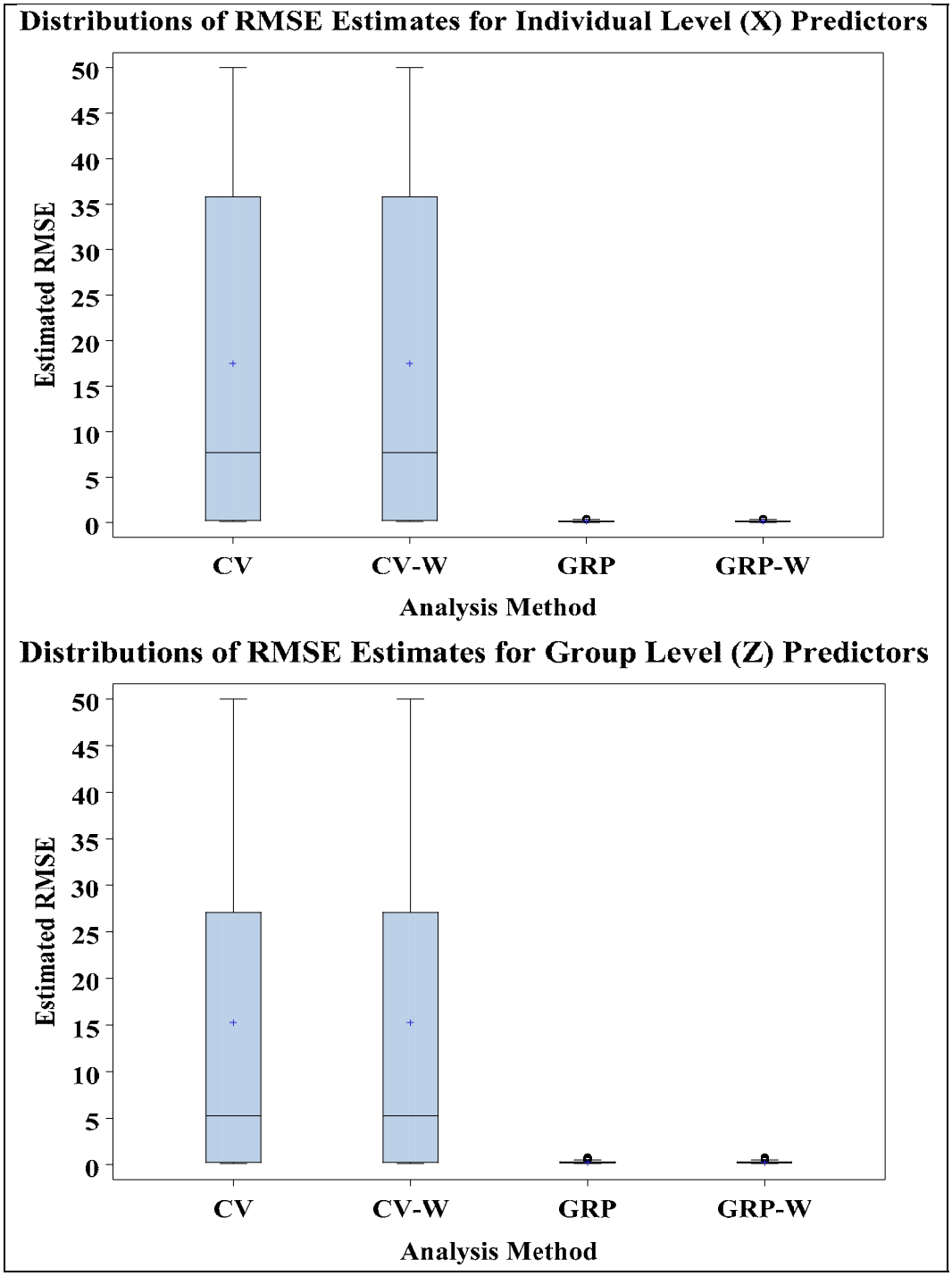

Figure 4: Distributions of Estimated RMSE Across Study Conditions (α = .05)

5. Method (Study 2)

To investigate the performance of the FIML approach, we generated data using the same method as in the first simulation. Data were generated for 25, 100, and 500 groups, with group sizes of 5-15 and 20-60, and an ICC of .20. We used two group level predictors and three and seven individual level predictors, with reliability levels of .70 and 1.00. Correlations between individual level predictors were set to .10, between the group level predictors to .20 and .40, and cross-level correlations were set to .30 and .40. We investigated effect sizes of zero (to estimate Type I error control) and 0.15 (to estimate power). The FIML estimation was implemented in Mplus Version 6 (Muthén & Muthén, 2007).

6. Results (Study 2)

6.1 Type I Error Control

The distributions of estimated Type I error rates across the simulation conditions are provided in Figure 5. In contrast to the CV and GRP approaches (with and without White's adjustment), the FIML approach resulted in notably elevated Type I error rates for tests of regressors at both levels. With the failure to control Type I error probability evidenced by the FIML approach, estimation of statistical power was not undertaken.

6.2 RMSE in Parameter Estimates

The distributions of RMSE in the parameter estimates are provided in Figure 6. The FIML approach resulted in large average values and large variability in the RMSE estimates. These values were larger than those provided by the CV and CV-W approaches. The use of GRP and GRP-W resulted in notably smaller values of RMSE, and consistently small values across the simulation conditions.

7. Results (Model Convergence)

Both the CV and FIML approaches to analyzing the simulated samples evidenced problems with model convergence in some conditions. Overall, the CV approach failed to converge in 7% of the samples (a rate consistent with that reported by Croon & van Veldhoven, 2007). A much larger problem with model convergence was seen with the FIML approach. When three Individual level predictors were used, only 44% of conditions converged with all samples and non-convergence rates reached as high as 10% of the samples. When the number of Individual level predictors was increased to seven, only 22% of conditions converged with all samples and non-convergence rates reached as high as 45% of the samples. As expected, the GRP approach converged with all samples simulated.

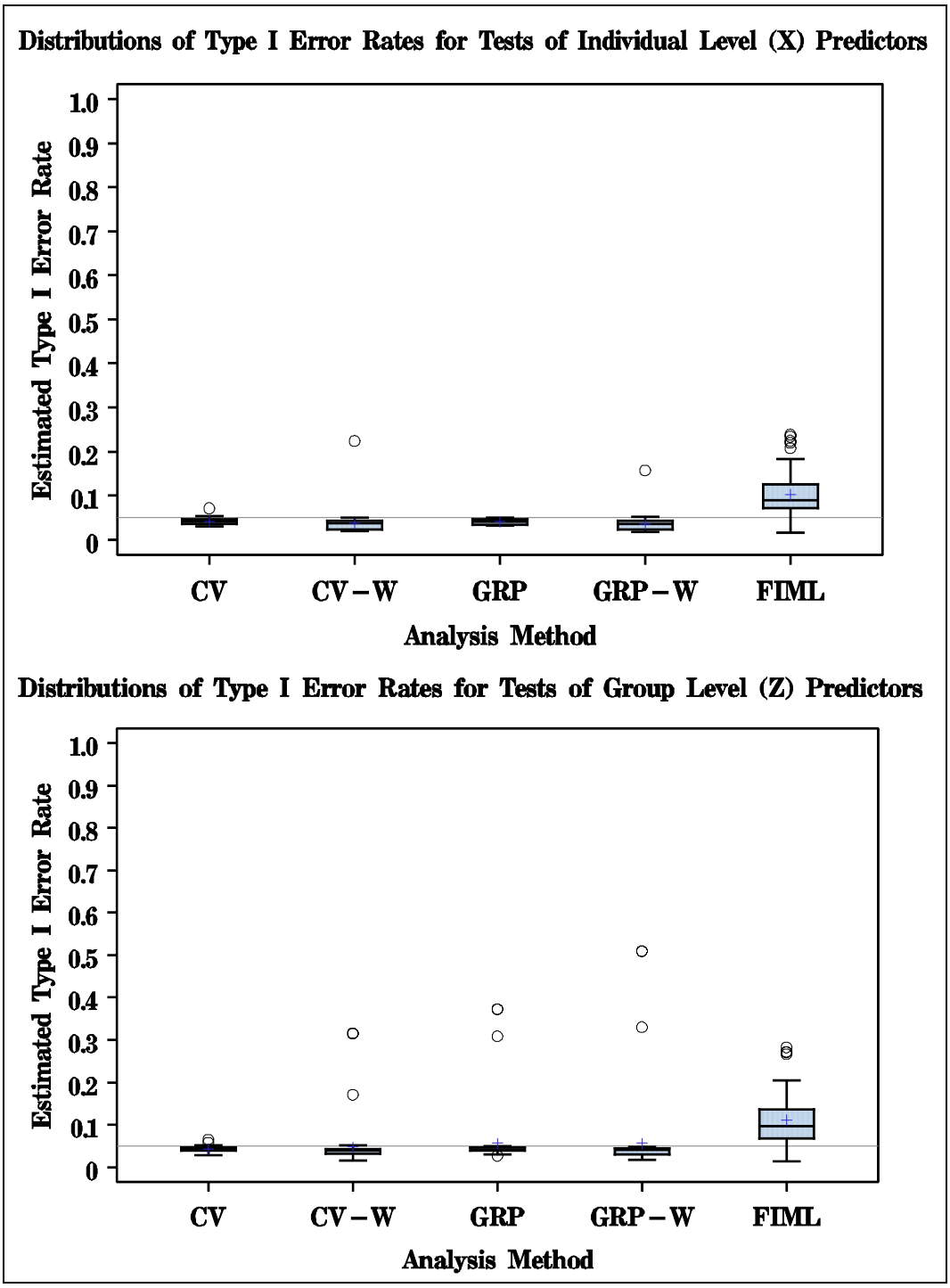

Figure 5: Distributions of Estimated Type I Error Rates Across Study Conditions (α = .05)

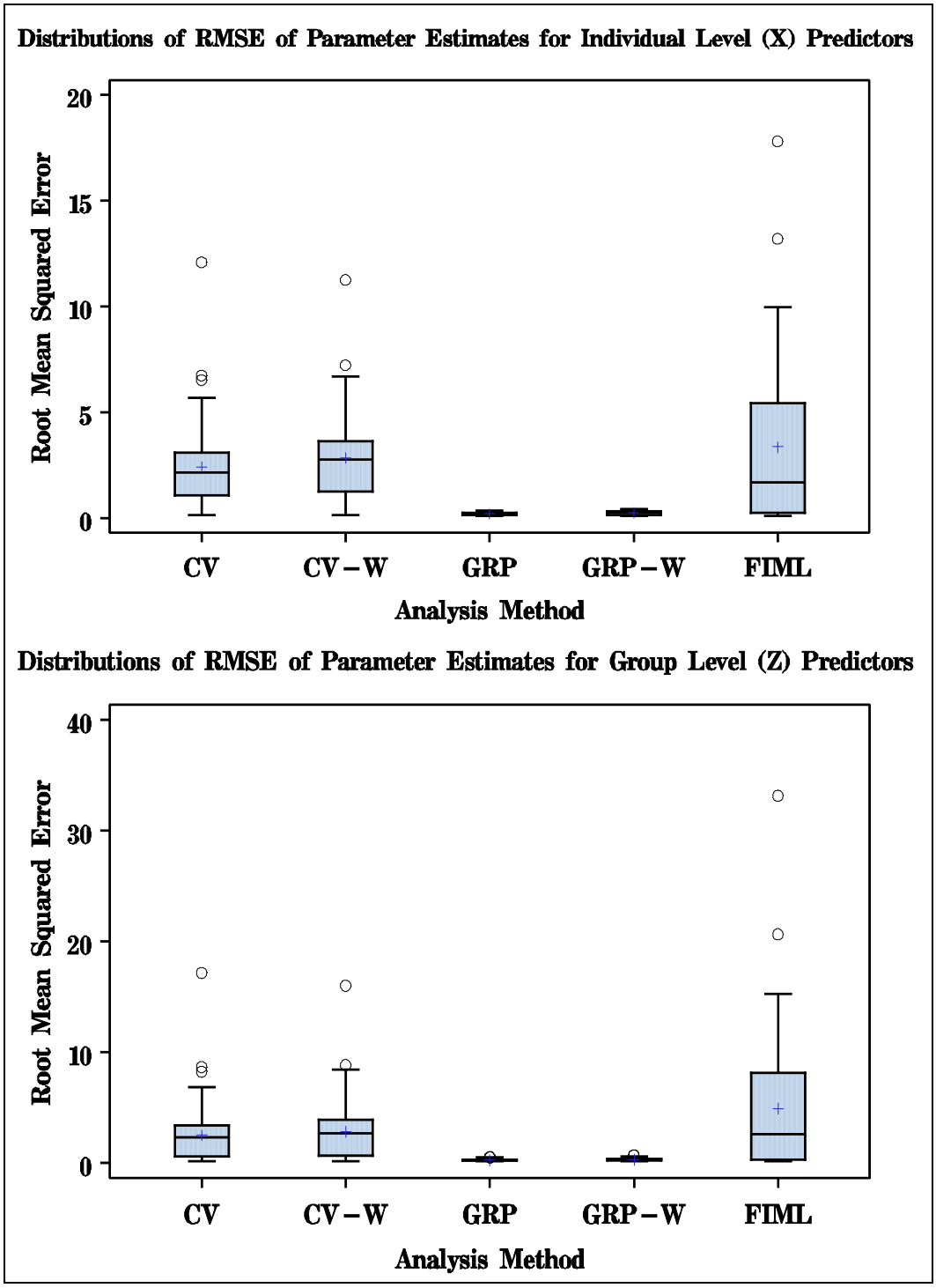

Figure 6: Distributions of Estimated RMSE Across Study Conditions (α = .05)

8. Conclusions and Recommendations

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. 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 and cross-level correlation that differs for individual and group level predictors. For individual level (X) predictors, as the number of groups increased, 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 the number of groups and the cross-level correlation increased, statistical power improves. For group level predictors using the CV approaches, statistical power decreases as the number of groups and the cross-level correlation increased. For both approaches, the power magnitude and impact of the interaction is more prominent when White's correction is utilized.

Our analyses of the FIML approach suggested by Lüdtke et al. (2008) yielded similar findings in that the method does not provide adequate Type I error control for the types of data conditions investigated in this study. Further, for conditions in which Type I error control was adequate, the FIML approach provided less statistical power for tests of group level predictors than that provided by the GRP-W analysis. Finally, our results showed that severe problems with non-convergence occurred with the FIML method as the number of individual level predictors increased unless the number of groups was 500 (even with 500 groups, up to 7% of the samples did not converge in some conditions). The reader is reminded that this research differs from the Lüdtke et al. research in two important ways. First, Lüdtke et al. examined an outcome variable measured at the individual level rather than at the group level. Secondly, Lüdtke et al. included only a simple model in their simulations (using one regressor at the individual level and one at the group level). The models investigated in the current study include multiple predictors at both levels with a variety of correlation patterns among them.

While 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 GRP-W and CV-W may temper our recommendations. 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 approach acceptable levels (power = .80) with little to no cross-level correlations and at least 100 groups, combined with White's correction. For predictors at the group level (Z), using the GRP approach in conjunction

with Whites 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.

In the context of analyzing multilevel data with the outcome at the group level, the general guidelines for selecting an analysis strategy and 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)

References

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.

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

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.

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**. (**2008). Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis, 29,* 60-87.

Hox, J. (2002). *Multilevel analysis: Techniques and application*. Mahwah, NJ. Lawrence Erlbaum.

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., Kashy, D.A., Mannetti, L., Pierro, A., & Livi, S. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology, 83*, 126-137.

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.

Maxwell, S.E., Delaney, H. D. & Dill, C.A. (1984). Another look at ANCOVA versus blocking, *Psychological Bulletin, 95*, 136-147.

Muthén, L. K. & Muthén, B. O. (2007. *Mplus user's guide*. Los Angeles: Muthén & Muthén.

Nunnally JC. (1970). *Introduction to psychological measurement.* New York: McGraw-Hill.

Nunally**,** J.C. **(**1978**).** *Psychometric Theory*. McGraw-Hill, New York.

O'Brien, R. M. (1990). Estimating the reliability of aggregate-level variables based on individual-level characteristics. *Sociological Methods and Research, 18*, 473-504.

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

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]. Carey, NC: SAS Institute, Inc.

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

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.