2-12-2010

Power and Bias in Hierarchical Linear Growth Models: More Measurements for Fewer People

Regine Haardoerfer
Georgia State University

ACCEPTANCE

This dissertation, POWER AND BIAS IN HIERARCHICAL LINEAR GROWTH MODELS: MORE MEASUREMENTS OF FEWER PEOPLE by REGINE HAARDOERFER, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chair, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

_____
Phill Gagné, Ph.D.
Committee Chair

_____
Paul A. Alberto, Ph.D.
Committee Member

_____
L. Juane Heflin, Ph.D.
Committee Member

_____
Frances A. McCarty, Ph.D.
Committee Member

_____
Date

_____
Sheryl A. Gowen, Ph.D.
Chair, Department of Educational Policy Studies

_____
R. W. Kamphaus, Ph.D.
Dean and Distinguished Research Professor
College of Education

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education's director of graduate studies and research, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____

Regine Haardoerfer

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Regine Haardoerfer
705 Vista Leaf Court
Roswell, GA 30075

The director of this dissertation is:

Dr. Phill Gagné
Department of Educational Policy Studies
College of Education
Georgia State University
Atlanta, GA 30303 – 3083

VITA

Regine Haardoerfer

ADDRESS: 705 Vista Leaf Court
Roswell, Georgia 30075

EDUCATION:

Ph.D.	2010	Georgia State University Educational Policy Studies: Research, Measurement, Statistics
M.Ed.	2005	Western Governors University Management and Innovation
2nd State Exam	1999	State of Bavaria Teaching Secondary Mathematics, Physics, and Computer Science
1st State Exam	1997	University of Erlangen-Nürnberg Teaching Secondary Mathematics and Physics

PROFESSIONAL EXPERIENCE:

2006–2010	Graduate Research Assistant Georgia State University, Atlanta, GA
2003-2005	Private Tutor
2000-2003	IB Teacher, 12th Grade Advisor, and Technology Coordinator Atlanta International School
1999-2000	IT Training and Support Atlanta International School

PROFESSIONAL SOCIETIES AND ORGANIZATIONS:

2006–present	American Educational Research Association
2007–present	The Association of Teacher Educators
2008-present	American Educational Studies Association

PRESENTATIONS AND PUBLICATIONS:

Davis, D. H., Gagné, P., Haardörfer, R., & Waugh, R. E. (2010, May). *Examining reading instruction for students with moderate intellectual disabilities using visual analysis and growth modeling*. Paper accepted for presentation at the annual convention of the Association for Behavioral Analysis International, San Antonio, TX.

Haardörfer, R., & Gagné, P. (2010, April). *Power and bias in hierarchical linear growth models:More measurements of fewer people*. Paper accepted for presentation at the annual meeting of the American Educational Research Association, Denver, CO.

Haardörfer, R., & Gagné, P. (in press). The use of randomization tests in single-subject research. *Focus on Autism and Other Developmental Disabilities*.

Haardörfer, R. (November 2009). International Baccalaureate in a Neoliberal Climate. Paper presented at the Annual Conference for the American Educational Studies Association. Savannah, GA.

Haardörfer, R. (2007). Teacher Autonomy. Paper and Performance presented at the Annual Conference for Curriculum and Pedagogy Group. Austin, TX.

Meyers, B., Swars, S., Schafer, N., Kavanagh, K., Haardörfer, R., Parrish, C., Jacobs, L., Matthews, C., and Taylor, S. (March 2008). Themes and Variations of Critical Friends Groups: A Contextual Look and Virtual Demonstration. Paper presented at the Annual Conference for the American Educational Research. New York, NY.

Meyers, B., Schafer, N., Parrish, C., Taylor, S., Swars, S., Kavanagh, K., Haardoerfer, R., Lick, B., Matthews, S., Jacobs, L. (February 2008). Themes and Variations of Critical Friends Groups: A Contextual Look and a Virtual Demonstration. Paper presented at the Annual Conference for the Association of Teacher Educators, New Orleans, LA.

Schafer, N., Kavanagh, K. Meyers, B., Swars, S., Czaplicki, K., and Haardörfer, R. (April 2009). Virtual Critical Friends Groups: A Vehicle for Professional Development and Supporting Beginning Teachers. Paper presented at the Annual Conference for the American Educational Research. San Diego, CA.

Schafer, N., Meyers, B., Swars, S., Haardörfer, R., Kavanagh, K., and Czaplicki, K. (April 2009). A Comparative Multicase Study Examining the Affordances and Constraints of Critical Friends Groups. Paper presented at the Annual Conference for the American Educational Research. San Diego, CA.

Schafer, N., Meyers, B., Swars, S., Kavanagh, K., Haardoerfer, R., Czaplicki, K.. and Matthews, C. (March 2009). Affordances and Constraints of Critical Friends Groups: Year Two Data of a Longitudinal Comparative Study and Virtual Demonstration. Paper presented at the Annual Conference for the Association of Teacher Educators, Dallas, TX.

Smith, M. E., Swars, S. L., Smith, S. Z., Hart, L. C., & Haardörfer, R. (2009). *A comparative longitudinal study of mathematics beliefs and knowledge in a changing elementary education program.* Paper presented at the 31st Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Atlanta, GA.

ABSTRACT

POWER AND BIAS IN HIERARCHICAL LINEAR GROWTH MODELS: MORE MEASUREMENTS OF FEWER PEOPLE

by
Regine Haardoerfer

Hierarchical Linear Modeling (HLM) sample size recommendations are mostly made with traditional group-design research in mind, as HLM as been used almost exclusively in group-design studies. Single-case research can benefit from utilizing hierarchical linear growth modeling, but sample size recommendations for growth modeling with HLM are scarce and generally do not consider the sample size combinations typical in single-case research. The purpose of this Monte Carlo simulation study was to expand sample size research in hierarchical linear growth modeling to suit single-case designs by testing larger level-1 sample sizes (N1), ranging from 10 to 80, and smaller level-2 sample sizes (N2), from 5 to 35, under the presence of autocorrelation to investigate bias and power. Estimates for the fixed effects were good for all tested sample-size combinations, irrespective of the strengths of the predictor-outcome correlations or the level of autocorrelation. Such low sample sizes, however, especially in the presence of autocorrelation, produced neither good estimates of the variances nor adequate power rates. Power rates were at least adequate for conditions in which N2 = 20 and N1 = 80 or N2 = 25 and N1 = 50 when the squared autocorrelation was .25.Conditions with lower autocorrelation provided adequate or high power for conditions with N2 = 15 and N1 =

50. In addition, conditions with high autocorrelation produced less than perfect power rates to detect the level-1 variance.

POWER AND BIAS IN HIERARCHICAL LINEAR GROWTH MODELS: MORE MEASUREMENTS OF FEWER PEOPLE

by
Regine Haardoerfer

A Dissertation

Presented in Partial Fulfillment of Requirements of the
Degree of
Doctor of Philosophy
in
Educational Policy Studies
in
the Department of Educational Policy Studies
in
the College of Education
Georgia State University

Atlanta, GA
2010

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

Chapter 1

Introduction

Single-case research aims to assess the changes in the behavior of participants. This is accomplished by investigating functional relations between the treatments (independent variable) and the behaviors (dependent variable). A functional relation is "a quasi-causative relation between the dependent and independent variables [that] exist[s] if the dependent variable systematically changes in the desired direction as a result of the introduction and manipulation of the independent variable" (Alberto & Troutman, 2009, p. 425). Repeated measures are used to investigate the impact of an intervention or treatment; that is, study participants are measured frequently on an outcome variable (behavior) under different conditions.

Single-case research can be categorized as time-series research, which is defined as

> a periodic measurement process on some group or individual and the introduction of an experimental change into this time series of measurement, the results of which are indicated by a discontinuity in the measurements recorded in the time series. (Campbell, Stanley, & Gage, 1966, p. 37)

Primarily, the research participant is compared to her- or himself; comparisons across individuals are important but generally secondary.

Single-case data sets have specific characteristics. The number of participants is low; often only a few individuals can be recruited for a study. Furthermore, the number of measurements per participant is higher than in traditional repeated measures designs,

where two or three waves are most common, but lower than in classical time-series research, which often produces more than 100 data points.

The traditional approach to analyzing single-case data is visual analysis. The data are graphed with time on the horizontal axis and the dependent variable on the vertical axis. Visual analysis investigates the presence and shape of a functional relation between the treatment and the outcome measure. The focus of the analysis is on the three main characteristics of single-case data: central location, trend, and variability (Franklin, Gorman, Beasley, & Allison, 1997). Central location can take on different definitions such as mean, mode, or median. Trend indicates a systematic, but not necessarily monotonic, increase or decrease of the outcome over time. Variability assesses the residuals after central location and trend have been taken into account.

While visual analysis is an established part of single-case data analysis, its critics debate several issues. Researchers indicate that interrater reliability might be a problem (Jones, Weinrott, & Vaught, 1978; Matyas & Greenwood, 1990). Furthermore, visual analysis is successful at keeping Type I errors low, as visual analysts are conservative in their assessment of a functional relation (Kazdin, 1982). That, however, leads to low power, especially for small and medium effect sizes.

One proposed approach to remedying these issues is the use of statistical analyses in addition to visual analysis. Several suggestions have been made over the last few decades, but more recently, efforts have increased due to changes in federal funding requirements. Since 2001, the No Child Left Behind (NCLB) Act ("No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425," 2002) requires studies to be “scientifically based research,” which according to Article 37, is defined as “research

that involves … objective procedures to obtain reliable and valid knowledge." ("No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425," 2002). This definition excludes studies from receiving federal funding if they rely solely on visual analysis.

Two schools of thought prevail on the use of statistical analyses for single-case data. Some researchers advocate non-parametric tests tailored to single-case data. Others suggest the use of well established group analyses.

Randomization tests are the most popular non-parametric tests that have been proposed and used in single-case research (Todman & Dugard, 2001). On the surface, they offer simple methods to determine statistical significance. They are, however, incompatible with single-case research philosophy and controversial due to unsubstantiated claims regarding their validity. Their low statistical power only exacerbates these problems (Haardörfer & Gagné, in press).

Most traditional statistical analyses such as t-tests and ANOVA focus on group comparisons and condense data into phase averages. The reduction of data into phase averages results in loss of information regarding individual responses. They are therefore incongruent with single-case researchers' interests (Barlow & Hersen, 1984). Furthermore, such analyses also are based on assumptions that are incongruent with certain characteristics of single-case data, such as the presence of autocorrelation (also called serial-dependency), which is a common occurrence when taking many measurements of the same people in a relatively short amount of time.

Although autocorrelation is an established part of time-series research (Yaffee & McGee, 2000), the literature on autocorrelation in single-case research is rich in

discussions about the presence and impact of autocorrelation. A lengthy discussion originated in the mid-1980s surrounding the question of whether autocorrelation even exists in single-case data (Busk & Marascuilo, 1988; Huitema, 1985, 1988; Sharpley & Alavosius, 1988; Sideridis & Greenwood, 1997; Suen, 1987; Suen & Ary, 1987). This exchange was started by Huitema's (1985) claim that autocorrelation is merely a myth. Several studies investigating the presence of autocorrelation in single-case data ensued, yielding diverse results. Sideridis and Greenwood (1997) suggest that autocorrelation is present in only 12% of the baselines of an extensive pool of single-case behavioral experiments. Bengali and Ottenbacher (1998) encountered autocorrelation more often in treatment than in baseline phases and with higher values. As time-series methodologists have pointed out for a long time, in research involving treatment that aims to change behavior, "serial dependency will coexist" (Sideridis & Greenwood, 1997, p. 290). Thus, regarding autocorrelation Sideridis and Greenwood appropriately urge researchers that "whenever statistical analyses are contemplated, its presence should always be examined" (p. 273).

Some researchers have suggested the use of regression analysis as a way to analyze single-case data (Allison & Gorman, 1993; Center, Skiba, & Casey, 1985). Regression analysis does not require as many data points as time-series analysis. One problem, however, is that regression is inappropriate for nested data. Single-case data are nested as soon as data are collected from more than one participant; measurements are nested within people.

Expanding on the idea of regression analysis, methodologists have investigated the possibility of using multilevel models to gain information across individuals without

losing the focus on the individual. Several researchers have suggested that hierarchical linear modeling (HLM) offers a viable option to single-case researchers (Jenson, Clark, Kircher, & Kristjansson, 2007; Lumpkin, Silverman, Weems, Markham, & Kurtines, 2002; Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003a, 2003b; Zucker, Schmid, McIntosh, Agostino, Selker, & Lau, 1997). Originally dealing with cross-sectional nested data (e.g., students in schools), HLM has been expanded to analyze repeated measures where measurements are nested in individuals. The first level models the growth of an individual, just as a single-case researcher displays a graph for each study participant. The second level of the model includes person-level predictors. A third level can be introduced, if individuals are nested in groups such as classrooms or therapy groups.

While single-case data fulfill the key requirements to use HLM analyses, the often small level-2 sample sizes present a challenge. The HLM literature does not offer recommendations addressing sample sizes in the range of those used by single-case researchers. While it is widely acknowledged that larger sample sizes on either level yield better estimates and higher power (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999), applied researchers have to consider financial and methodological constraints. Thus, detailed knowledge regarding sufficient sample sizes would be invaluable. The present study addressed this issue by investigating sample size combinations that are realistic for single-case research designs featuring many measurements for relatively fewer people than typically recommended for group designs.

Chapter 2

Literature Review

The majority of HLM power recommendations address cross-sectional studies of different degrees of model complexity. Recommendations based on analytical approaches exist for two-level models without level-2 predictors. Raudenbush (1997) provides formulae to estimate the optimal level-1 (N1) and level-2 (N2) sample size for such models, including cost factors connected to increasing sample size on either level. A simulation study by Mok (1995) attempts to address the same questions. Her design includes a symmetrical use of sample sizes on both levels ranging from 5 to 150, resulting in total sample sizes ranging from 25 to 22500. Mok's results indicate that increasing N2 has a greater positive impact on bias and power than increasing N1. Her suggestions call for a total sample size of 3500 for an intra-class correlation below .15 to ensure sufficiently low bias in the estimates. Browne and Draper (2000), however, report that estimates were close to being unbiased for a total sample size of 216. They acknowledge that power was not acceptable in this case, but was adequate for a total sample size of 864. The different results might be attributed to the researchers' choices in values for the fixed and random effects.

Beyond the models without level-2 predictors, several authors offer analytical considerations or formulae for models that include dichotomous predictors at level 2 (O'Connell & McCoach, 2008; Raudenbush, 1997). O'Connell and McCoach offer recommendations for two- and three-level random-intercept random-slope models. For these models, the authors provide formulae to estimate sample sizes. They find that increasing the number of clusters increases the power to detect the fixed and

random effects, as does including a cluster-level predictor. For more specific recommendations, they refer to available power analysis software programs that require researchers to provide estimates of effect sizes, intra-class correlation, and proportion of variance explained by the level-2 predictor. Expanding on Raudenbush's (1997) earlier work, Raudenbush and Liu (2000) conclude that Mok's findings that the number of sites has greater effect than the number of participants per site also holds true for models with dichotomous predictors at level 2.

Regarding models with one or more non-dichotomous predictors at the second level, advice on minimum sample sizes is less plentiful. Even established textbook authors offer little in the way of concrete recommendations. Raudenbush and Bryk (2002) address the issue of sample size toward the end of the book and only in very general terms. Kreft and De Leeuw (1998) consider group sizes of 10 to be small and 100 to be large. In addition, they cite three unpublished dissertations that provide research regarding sample size recommendations. The authors conclude that the total number of observations is key for level-1 parameter estimates, while the number of groups has a clear impact on the power of level-2 estimates. They suggest that both sample sizes should be larger than 30 to detect a cross-level interaction. In addition, they state that having a larger level-2 sample size has a greater impact on power than a larger level-1 sample size, holding the total number of measurements constant. In addition, they contend that a level-2 sample size of 150 leads to a high power (0.90) for a low level-1 sample size of 5. The authors also mention, however, that a downward bias in variance estimates is present in level-2 samples of less than 300 for all tested level-1 sample sizes. The authors conclude that effect size as well as intra-

class correlation play a role in all of these recommendations but do not make mention of how these factors could or should be taken into account. Snijders and Bosker (1993) approach the topic analytically and provide the reader with approximation formulae to estimate optimal sample sizes. Specifically, Snijders and Bosker recommend N1 to be larger than 10. The researcher, however, needs to provide a lot of information regarding the data: variance components, means, and covariance matrices need to be known or estimated to be able to calculate more specific minimum sample sizes.

Using simulation methods, Maas and Hox (2005) tested scenarios with N2 being 30, 50, or 100, while N1 was 5, 30, or 50. They crossed these possibilities with three values for intra-class correlations. Their results supported their hypothesis that any of the given sample size combinations lead to sufficient power. Two more general Monte Carlo studies regarding cross-sectional HLM have been conducted recently by Gagné and Estes (2009a, 2009b). In their comprehensive studies, they tested models with either one predictor at each level or two in one level and one at the other level. Furthermore, they included a range of predictor-criterion values as well as presence or absence of cross-level interaction as parameters, thus producing a controlled correlational structure. Their studies further support the finding that an increase of sample size at either level increases power but an increase in level-2 sample size has greater impact. Gagné and Estes (2009a) conclude that level-2 sample sizes smaller than 35 can yield acceptable power with at least 10 measurements per group.

The literature regarding minimum sample size recommendation for hierarchical linear growth models is not plentiful either. The textbook on applied longitudinal data analysis by Singer and Willet (2003) mentions sample sizes only in regard to the minimum number of measurements per participant. Two groups of researchers (Raudenbush & Liu, 2001; Zhang & Wang, 2009) offer advice regarding sample size recommendations for data with independent error structures. Zhang and Wang (2009) offer SAS macros to calculate the power given N1, N2, effect size, and number of participants for linear and quadratic growth models. While the macros can account for systematic attrition of study participants, they do not include level-2 predictors or non-independent error covariance structures. Sample conditions indicate that an effect size of 0.2 for the slope might require more than 300 participants when measured three times and still more than 200 when six measurements are taken per person. The authors also show power curves dependent on slope effect size which illustrate that power increases quickly with an increase in effect size.

Raudenbush and Liu (2001) focused on models with a dichotomous level-2 predictor with independent error variance-covariance structures. Like O'Connell and McCoach (2008), their theoretical work leads them to the general conclusion that an increase in N1 or N2 increases power. Their recommendations are two-fold: increasing N1 is best when the degree of the polynomial is high and there is considerable within-person variance; increasing N2 is best when between-person heterogeneity is large. Similar to Zhang and Wang (2009), Raudenbush and Liu's level-2 sample size recommendations are quite large, ranging from 238 to 800 for an effect size of 0.40 depending on frequency of observation.

Hedeker, Gibbons, and Waternaux (1999) also offer analytical solutions for minimum sample sizes. They focus on non-independent error variance-covariance structures including compound symmetry, random effect, and autocorrelation. Their formulae allow the researcher to estimate a minimum level-2 sample size to reach power of .80 if she knows the level-2 sample size, the effect size, all predictor-outcome correlations, and the within- and between-group variances. The authors furthermore offer a table with some pre-calculated minimum sample sizes for level-2 for 4, 6, and 8 measurements, each instance pertaining to a power of .80. The authors estimate that for 4 waves of data with an autocorrelation of .3, adequate power to detect a small linear effect necessitates 758 participants. The level-2 sample size recommendation decreases to 48 with large effect sizes and decreases even further for larger autocorrelations. Interestingly, in the presence of high autocorrelation, an increase in level-1 sample size calls for an increase in level-2 sample size to achieve adequate power.

While the aforementioned researchers focused exclusively on power, estimation bias is of concern as well. As HLM uses maximum likelihood estimations, any model misspecification can cause bias in the estimation of fixed and random effects (White, 1982). Ferron, Daily, and Yi (2002) found that assuming independence of errors instead of allowing for a nonzero autocorrelation introduces only a small bias in the estimates of the fixed effects but inflated the error variances much more. The biases diminish slightly, however, with an increase in level-2 sample size. The authors caution that the biases might be much larger in more complex

models than theirs. Kwok, West, and Green (2007) support these findings and add that the inflation of estimates of the error variances impacts power negatively.

Two recent studies include autocorrelation in their investigation of bias and power in using HLM for single-case designs (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Jenson et al., 2007). Both focus on the most basic building block of single subject design (Alberto & Troutman, 2009), the AB design. The model includes one level-1 predictor, a dummy variable, to indicate the beginning of the treatment phase. No level-2 predictors were considered. Ferron et al. (2009) indicate that power was high for very small sample sizes of 4 measurements and 10 participants, though lower in the presence of autocorrelation. Jenson, Clark, Kircher, and Kristjansson (2007), however, report that power was not sufficient for 15 participants, with 5 baseline and 10 treatment measurements, with or without autocorrelation. Their simulations suggest that even having 10 baseline and 20 treatment measures for 15 people does not yield adequate power for autocorrelations of .40 and .80. The discrepancies are likely due to the researchers' different choices for the value of the fixed and random effects.

In conclusion, sample size recommendations are fairly common for cross-sectional HLM, scarce for growth modeling including complex models, and emerging for single-case models. The purpose of the present study was to expand sample size recommendations in hierarchical growth modeling to larger numbers of measurements and smaller numbers of participants. It focused on a two-level model with one continuous level-1 predictor and one continuous level-2 predictor. The impact of autocorrelation was investigated as well.

Chapter 3

Method

This is a Monte Carlo simulation with five independent variables: the level-1 sample size (i.e., number of measurements per participant), the level-2 sample size (i.e., the number of participants), the magnitude of the autocorrelation factor of the level-1 error with lag 1, the magnitude of the interpredictor correlation, and the magnitude of the correlation between Time and the outcome variable Y. Data were simulated for a hierarchical model reflecting linear growth with one level-2 predictor for the intercept and the same level-2 predictor for the level-1 slope, as illustrated in the following equations for the level-1, level-2, and combined models

$$Y_{ti}=\pi_{0i}+\pi_{1ij}(TIME_{ti})+e_{ti}$$

$$\pi_{0i}=\beta_{00}+\beta_{01}X_i+r_{0i}$$
$$\pi_{1i}=\beta_{10}+\beta_{11}X_i+r_{1i}$$

or

$$Y_{ti}=\beta_{00}+\beta_{01}(X_i)+\beta_{10}(TIME_{ti})+\beta_{11}(X_i)(TIME_{ti})+r_{0i}+r_{1i}(TIME_{ti})+e_{ti}.$$

Single-case data do not meet the assumption that e_{ti} is normally distributed with a mean of 0 and a variance of σ_e^2. Instead, single-case researchers have to assume that the level-1 errors are autocorrelated. Most common is a lag of one, meaning that any consecutive error is autocorrelated with the error of the previous measurement,

$$e_{ti}=\rho e_{(t-1)i}+\nu_t.$$

In this formula, ρ represents the constant autocorrelation factor with an absolute value less than1 and ν_t is normally distributed with mean 0 and variance σ_e^2. Thus,

$$\sigma_e^2 = \frac{\sigma_v^2}{1-\rho^2},$$

where σ_e^2 denotes the apparent, and inflated, error variance as it would be detected if the analysis were not to look for the possibility of autocorrelation of the level-1 errors; σ_v^2 is the actual level-1 error variance in the data.

SAS 9.1 (SAS Institute, 2004) was used to generate the data and to conduct all calculations and analyses. For each combination of parameters, 1,000 replications were conducted, with the data being simulated in IML and then analyzed using the following PROC MIXED routine (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006; Singer, 1998):

```
PROC MIXED DATA=CompleteDataSet COVTEST NOCLPRINT
NOITPRINT NOINFO IC;
CLASS ID Wave;
MODEL Y = Time X TimeX /DDFM=BW CL NOTEST;
RANDOM Intercept Time /SUB=ID TYPE=UN;
REPEATED Wave /SUB=ID TYPE=AR(1);
RUN;
```

To simplify the situation, time points were spaced equally. Thus, the level-1 sample size was directly proportional to the duration of the "study." For ease of interpretation, the spacing was 1. In addition, the level-2 predictor was grand-mean centered, and Time was simulated such that Time = 0 at the midpoint of the measurement occasions.

A broad range of sample sizes pertinent to single-case research was investigated. The level-1 sample size took on seven different values: 10, 15, 20, 30, 40, 50 and 80. The level-2 sample sizes, reflecting the number of participants, were 5, 10, 15, 20, 25, 30, and 35. This lead to 36 combinations of sample sizes. The

autocorrelation factors were chosen so that their squares were .05, .10, .15, .20, and .25. Time and the outcome variable were correlated at .2, .3, .4, and .6. The correlation between the level-2 predictor and the outcome variable was set to .2 and .3. The correlation between the cross-level interaction and the outcome variable was fixed at .3, and the correlation between Time and the level-2 predictor was fixed at 0. These parameters lead to 36 x 6 x 4 x 2 conditions and thus to 1728 cells.

Power for each set of parameters was calculated as the percentage of betas that are identified as statistically significant. Power results ranging from 0.80 to 0.90 were considered acceptable, while power greater than 0.90 was considered high. Relative bias in parameter estimates was considered low if it is below 5%, and relative bias less than 10% in the standard errors was considered low (Hoogland & Boomsma, 1998), and it was calculated as

$$\text{bias} = \frac{\text{parameter estimate - parameter}}{\text{parameter}} * 100.$$

Chapter 4

Results

The results regarding the biases of the estimates are presented in color-coded tables. The color green is used for magnitudes below 5%. Yellow illustrates biases with magnitudes between 5% and 10%. Orange signifies conditions in which the magnitudes of the biases exceeded 10%.

For all power tables, the cells shaded in green indicate high power that is power of at least .90 but less than 1. Perfect power, a power of 1, is marked dark green. Yellow signifies conditions in which power was adequate, defined as at least .80 and below .90. Any cells marked indicate inadequate power, or values below .80.

Conditions With Zero Autocorrelation

Biases. In general, the estimates of the fixed and random values were neither dependent upon the strength of the correlation between Time and the outcome variable nor between X and Y for any of the conditions tested. The variations in biases of the estimates when autocorrelation had been set to 0 only varied due to differences in sample size combinations. As expected, an increase in either level-1 or level-2 sample size improved the accuracy of the estimates of the fixed and the random effects.

For the conditions tested with autocorrelation of errors set at 0, the fixed effects (β_{00}, β_{01}, β_{10}, and β_{11}) were all estimated with the absolute values of the biases below 5% when N2 was at least 10, independent of the strength of the correlation between the predictors. The magnitudes of 9 of the 192 biases were slightly above 5% when N2 = 5 with none exceeding 6.63%. No clear pattern, however, was discernable.

For the conditions examined, the biases for the level-2 error variances (τ_{00} and τ_{11}) were below 5% for sample size combinations with a total sample size GrandN (N1xN2) larger than 150. The biases were especially large for conditions with a level-2 sample size of 5 (Tables 1 and 2). As level-1 sample sizes increased, however, the biases decreased quickly for both level-2 variances. With N2 = 5 and N1 = 50, biases were below or around 5%. At a level-1 sample size of 80, the magnitudes of the bias were all below 5%.

Table 1

Biases of Estimating τ_{00} When N2 = 5

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	25.8439	15.3692	9.3136	8.6327	5.1587	3.6403
	.3	28.6305	16.4368	13.3016	6.3959	4.9817	2.1879
	.4	26.9261	15.7487	13.7419	7.7229	3.1964	1.4446
	.6	32.3837	12.9452	12.303	9.3636	2.7376	-0.7201
.3	.2	30.5126	17.8394	14.4369	5.6151	1.312	4.0045
	.3	28.0941	12.9102	10.9528	6.0016	7.732	2.777
	.4	26.9775	18.3079	7.1991	4.482	1.3091	0.1018
	.6	29.0606	16.4689	7.8662	7.5697	0.6275	2.1589

Table 2

Biases of Estimating τ_{11} When N2 = 5

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	32.8617	18.8305	12.5967	7.6518	4.8994	2.919
	.3	29.86	14.6556	13.929	9.8568	4.0105	0.3235
	.4	30.0971	13.758	13.0104	4.916	0.7321	2.9474
	.6	37.9955	17.9014	13.193	3.2803	3.5469	0.9802
.3	.2	36.0297	18.146	10.1988	5.9091	3.0259	1.8507
	.3	34.8824	13.5734	10.1184	6.7328	5.0199	3.2689
	.4	33.8035	18.8711	11.3961	5.3816	-0.3078	3.98
	.6	29.8537	21.3782	15.0326	10.9989	2.0832	2.9362

With N1 = 10 and N2 = 10, about half of the bias values were above 5%. In addition, many biases for (10, 15)[1] or (15, 10) as a sample size combination were between 5% and 10%. Furthermore, neither an increase in the correlation between Time and the outcome variable nor an increase in the correlation between X and the outcome variable seemed to impact the variance estimation of the level-2 variances.

The level-1 error variances were all estimated with biases below 5% for all conditions tested for which the autocorrelation was 0. The strength of the predictor-outcome variables had no influence on the estimates. The sample sizes tested were all sufficiently large to produce good estimates of the level-1 variance.

[1] The values in parentheses indicate the values for the level-1 sample size and level-2 sample size according to their positions, (N1, N2).

Power. In all but two of the tested conditions with zero autocorrelation, the power to detect β_{00} was perfect, that is 100%. When N2 was 5 and N1 was 30, power decreased to 99.9% under the lowest correlation condition. The same result applied to the conditions in which both predictor-outcome correlations were set at .3 and the sample size combination was (5, 30). Power of the other fixed effects varied depending on the sample size combinations as well as the predictor-outcome correlations. Thus, they will be discussed separately in more detail.

Power to detect β_{10}. In general, power to detect β_{10} was similar for the two tested correlations between X and Y, with power values being slightly larger for the larger X-Y correlation. As expected, the value of the correlation between Time and Y had greater influence on the power to detect β_{10}. For the lowest level-2 sample size of 5, power was only adequate for the highest correlation between Time and the outcome variable (**Table** 3). When ρ_{TimeY} was set to .6, however, sample sizes of 15 and above yielded power greater than .8. Notably, when N2 was only 5 and the highest correlation combination was used, taking 80 measurements yielded power above .9.

Table 3

Power to Detect β_{10} When N2 = 5

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	15.70	19.40	21.90	25.50	25.60	29.90
	.3	27.80	32.90	33.80	39.30	41.20	46.30
	.4	44.00	45.50	49.10	53.30	59.20	65.70
	.6	72.60	78.30	80.70	82.90	86.60	88.90
.3	.2	16.70	19.00	22.00	23.10	29.10	32.80
	.3	29.70	34.20	36.90	40.60	43.50	48.90
	.4	42.40	48.80	53.90	56.80	61.80	69.00
	.6	77.20	81.40	83.40	84.20	87.60	92.10

Level-2 sample sizes of 10 and 15 produced high power rates when the correlation between Time and the outcome variable was large. Specifically, with N2 = 10 and ρ_{TimeY} = .6, power was above 90% for all tested level-1 sample sizes and X-Y correlations (Table 4). Independent of ρ_{XY}, when ρ_{TimeY} was .4, most conditions showed adequate or high power; when ρ_{TimeY} was .2 or .3, no conditions produced adequate power to detect β_{10}. Power rates increased when N2 was 15. Almost all conditions with ρ_{TimeY} of at least .4 produced high power (Table 5).

Table 4

Power to Detect β_{10} When N2 = 10

ρ_{XY}	ρ_{TimeY}	Level-1 Sample Size					
		10	15	20	30	50	80
.2	.2	28.30	32.40	33.20	34.30	39.90	46.20
	.3	50.20	55.90	57.00	64.40	66.40	73.50
	.4	72.90	76.00	79.40	83.40	85.50	89.40
	.6	97.30	97.80	98.30	98.30	98.60	99.70
.3	.2	31.40	31.00	33.20	36.70	38.20	46.90
	.3	53.10	57.60	59.70	62.70	66.50	76.90
	.4	74.90	81.10	82.50	84.80	86.70	92.00
	.6	97.60	98.50	98.50	99.50	99.40	99.60

Table 5

Power to Detect β_{10} when N2 = 15

ρ_{XY}	ρ_{TimeY}	Level-1 Sample Size					
		10	15	20	30	50	80
.2	.2	37.90	42.60	44.80	45.80	53.10	61.40
	.3	67.80	72.30	75.40	77.70	82.10	87.10
	.4	89.00	91.10	92.30	94.90	94.70	98.30
	.6	99.50	99.80	99.80	100.00	100.00	100.00
.3	.2	39.80	43.30	44.20	47.80	54.10	62.40
	.3	70.50	75.10	77.40	80.50	83.70	89.00
	.4	90.80	92.20	93.60	96.10	96.10	98.10
	.6	100.00	100.00	100.00	100.00	99.90	100.00

With N2 = 20 and ρ_{TimeY} = .6, almost all tested level-1 sample sizes yielded perfect power. Furthermore, when ρ_{TimeY} was .4, power was always above 90%. Even for ρ_{TimeY} = .3, all but the condition with the lowest values for N1 and ρ_{XY} had adequate or high power. None of the conditions with ρ_{TimeY} = .2 had adequate power with 20 participants.

Within a set of predictor-outcome correlations, power increased slowly with an increase in N2 (Table 6). This was especially pronounced for ρ_{TimeY} = .2. For conditions in which N2 $\geq$ 25, N1 $\geq$ 15, and $\rho_{TimeY} \geq$ .3 power was high. When only 10 measurements were taken per person, power was still adequate.

Table 6

Power to Detect β_{10} When N2 = 25

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	55.10	60.30	64.60	65.90	71.80	80.30
	.3	86.90	90.90	90.90	94.50	95.80	97.90
	.4	98.80	98.80	99.30	99.50	99.60	99.90
	.6	100.00	100.00	100.00	100.00	100.00	100.00
.3	.2	60.00	62.30	66.10	69.30	74.50	82.10
	.3	88.40	92.90	92.60	95.50	97.70	98.60
	.4	98.30	98.80	99.70	99.40	100.00	100.00
	.6	100.00	100.00	100.00	100.00	100.00	100.00

Even for the highest tested level-2 sample size of 35 (Table 7), power was not adequate for conditions where ρ_{TimeY} was .2 and N1 was less than 30. It was, however high for all other conditions. Most conditions with ρ_{TimeY} being .4 or .6 had perfect power.

Table 7

Power to Detect β_{10} When N2 = 35

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	69.60	74.40	78.10	81.60	84.40	92.30
	.3	95.70	97.40	97.30	98.70	99.10	99.70
	.4	99.90	99.80	100.00	99.80	100.00	100.00
	.6	100.00	100.00	100.00	100.00	100.00	100.00
.3	.2	72.40	74.40	79.30	83.10	86.00	92.60
	.3	96.30	97.90	97.40	99.00	99.60	99.70
	.4	100.00	100.00	100.00	100.00	100.00	100.00
	.6	100.00	100.00	100.00	100.00	100.00	100.00

Overall, the correlation between Time and the outcome variable had great influence on power to detect β_{10}. This was expected as β_{10} is the fixed effect that models the level-1 predictor's direct influence on the outcome variable. The magnitude of ρ_{XY} still had some impact on the power to detect β_{10}.

Power to detect β_{01}. Within any correlation combination, the power to detect β_{01} increased with an increase in either level-1 or level-2 sample size. It also increased with an increase in the correlation between the level-2 predictor X and the outcome variable Y, that is, the correlation associated with β_{01}. Furthermore, an increase in the correlation between the Time and the outcome variable increased the power of β_{01} as well. This increase, however, was not as strong as the one related to the increase in ρ_{XY}.

Power to detect β_{01} under conditions in which N2 = 5 was extremely low, ranging from 17.7% for the lowest level-1 sample size and lowest correlation combination to a still low 65.1% for the highest level-1 sample size and highest correlation combination. Additionally, when the correlation between the level-2 predictor and the outcome variable was set to .2, power was inadequate for all conditions with a level-2 sample size of either 10 or 15. For a correlation between X and Y of .3, power was at least adequate for more than half of the tested conditions, with values increasing with an increase in level-1 sample size or an increase in the correlation between Time and the outcome variable. In conditions when 80 measurements were simulated for 15 participants, power was high at all values tested of the correlation between Time and Y. Also, when the Time-Y correlation was set to .6, taking at least 15 measurements produced high power. The specific power results for N1 of 10 and 15 are presented in Table 8 and Table 9, respectively.

Table 8

Power to Detect β_{01} When N2 = 10

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	30.1	32.5	34.7	36.3	38.2	45.5
	.3	29.2	36.3	35.7	38.5	41.7	47.9
	.4	30.4	34.5	36.0	37.4	44.5	49.2
	.6	39.1	44.6	45.8	49.3	56.8	62.8
.3	.2	51.3	55.9	63.3	63.8	68.8	77.2
	.3	54.2	59.4	62.6	66.3	70.4	79.1
	.4	58.8	64.6	66.0	70.0	77.5	83.7
	.6	73.1	78.8	81.8	82.1	88.2	92.2

Table 9

Power to Detect β_{01} When N2 = 15

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	38.8	40.7	44.2	48.2	52.4	59.6
	.3	40.4	45.0	47.1	50.5	57.1	65.0
	.4	42.6	47.0	49.4	54.6	57.5	68.1
	.6	54.2	58.3	63.7	66.4	70.0	78.9
.3	.2	70.5	75.3	76.4	81.0	85.9	92.5
	.3	71.6	78.3	81.8	84.8	88.3	92.8
	.4	76.8	81.3	83.7	88.4	91.1	95.8
	.6	89.4	91.7	93.0	95.6	96.3	98.7

For a level-2 sample size of 20 (Table 10), power was never high when ρ_{XY} was .2 and ρ_{TimeY} was .6. It reached adequate values for level-1 sample sizes of 50 and 80. Power increased substantially when ρ_{XY} was increased to .3. All conditions in which ρ_{XY} = .3 with level-2 sample sizes of at least 20 had adequate or high power.

Table 10

Power to Detect β_{01} When N2 = 20

ρ_{XY}	ρ_{TimeY}	Level-1 Sample Size					
		10	15	20	30	50	80
.2	.2	47.5	52.6	55.5	59.4	64.5	73.7
	.3	50.9	54.1	57.3	60.8	66.5	75.6
	.4	52.6	58.5	62.4	65.5	70.5	79.8
	.6	64.5	72.6	73.4	78.2	84.4	88.6
.3	.2	83.7	86.3	89.0	91.2	94.7	97.1
	.3	83.8	88.4	89.8	92.4	95.3	98.6
	.4	88.2	91.0	91.2	94.8	96.8	99.1
	.6	94.9	97.1	98.4	98.9	99.7	99.9

When N2 was 25, power increased such that it was at least adequate for N1 = 80 and ρ_{XY} = .2 for all tested Time-outcome correlations. In the case of a high correlation of .6 between Time and Y and the low correlation of .2 between X and Y, power was adequate starting at a level-1 sample size of 15 and high when N1 was at least 50. When ρ_{XY} was increased to .3, all but the lowest sample size combination had high power. For N1 of either 50 or 80 and the correlation between Time and Y being at least .4, power was perfect (Table 11).

Table 11

Power to Detect β_{01} When N2 = 25

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	53.7	61.8	61.8	69.8	71.8	82.1
	.3	60.7	64.9	68	73.3	76.7	84.0
	.4	63.0	67.7	70.8	74.0	80.1	87.0
	.6	76.7	80.9	83.2	87.6	91.7	95.1
.3	.2	89.0	92.1	94.9	96.6	98.2	99.0
	.3	90.9	94.4	95.1	97.7	97.2	99.8
	.4	94.3	95.9	96.4	98.5	99.4	99.6
	.6	99.2	99.4	99.5	99.8	100	100

Table 12

Power to Detect β_{01} When N2 = 30

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	65.5	69.4	73.5	75.8	82.2	87.5
	.3	64.1	71.0	76.4	80.0	86.5	89.9
	.4	71.5	76.3	78.5	82.4	87.5	93.2
	.6	83.7	88.4	90.7	93.1	95.5	98.4
.3	.2	95.1	95.5	97.1	97.2	99.1	99.9
	.3	95.8	97.4	98.7	98.2	99.7	99.9
	.4	97.9	98.9	98.8	99.6	99.6	100
	.6	99.6	99.7	100	100	100	100

For a level-2 sample size of 30 or 35, all Time-Y correlations produced high or perfect power when $\rho_{XY} = .3$. For the conditions when ρ_{XY} was .2, power increased further with increasing level-1 sample size and ρ_{TimeY}. When N2 was 35, all conditions with N1 being at least 20 produced adequate or high power. For a Time-outcome correlation of .6, all tested values of N1 produced adequate or high power (Table 13).

Table 13

Power to Detect β_{01} When N2 = 35

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	71.5	77.1	80.3	82.0	88.5	92.9
	.3	75.6	78.0	81.2	84.0	89.6	94.2
	.4	78.1	81.0	85.6	86.8	91.9	95.8
	.6	87.5	94.2	94.2	96.1	97.1	99.5
.3	.2	97.9	98.2	98.7	99.6	99.8	100
	.3	97.8	99.0	98.9	99.5	100	100
	.4	98.5	99.1	99.5	99.7	99.9	100
	.6	99.9	100	100	100	100	100

Overall, the correlation between X and the outcome variable had great influence on power to detect β_{01}. This was expected as β_{01} is the fixed effect that models the level-2 predictor's direct influence on the outcome variable. When ρ_{XY} was .3, all conditions with a level-2 sample size of 20 or higher produced adequate or high power. In conditions with the lower correlation between the level-2 predictor and Y, only larger values of N1 and N2 produced high power.

Power to detect β_{11}. In general, the correlation between Time and the outcome variable (ρ_{TimeY}) had more impact on the power to detect β_{11} than the correlation between X and Y. Specifically, power was inadequate for all conditions tested where N2 was 5. Even for the highest correlation combination ($\rho_{TimeY} = .6$ and $\rho_{XY} = .3$), power to detect β_{11} was only 62%. When N2 was doubled to 10 (Table 14), only two conditions produced adequate power ($\rho_{TimeY} = .6$ and N1 > 50). When the level-2 sample size was increased to 15, power increased appreciably (Table 15). All conditions with a Time-Y correlation of at least .6 or at least 50 measurement occasions produced power rates above 80%.

This trend of increasing power continued with the increase of the level-2 sample size. For those conditions where N2 was 20, all but two conditions produced at least adequate power; for half of them, power was high. Furthermore, 42 conditions with N2 = 25 produced high power; the other 6 conditions produced adequate power (Table 16).

When N2 ≥ 30 power rates to detect the interaction effect were all high; more specifically, all power rates were above 94.9 %. Seven of the conditions with the highest number of measurements per person and higher predictor-outcome correlations reached perfect power (Table 17).

Table 14

Power to Detect β_{11} When N2 = 10

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	47.70	54.40	56.30	56.90	63.10	71.10
	.3	50.80	56.70	57.10	61.00	64.30	73.20
	.4	52.80	58.70	60.10	67.60	70.00	75.10
	.6	64.60	70.40	72.60	75.20	80.00	84.90
.3	.2	52.40	55.30	58.20	61.60	66.30	72.30
	.3	50.90	59.20	60.40	64.40	65.00	76.00
	.4	58.00	62.10	62.20	67.10	71.40	77.30
	.6	66.80	72.80	75.80	77.30	82.10	88.00

Table 15

Power to Detect β_{11} When N2 = 15

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	66.30	68.90	71.40	74.70	80.20	85.80
	.3	66.90	72.70	74.10	78.50	82.30	87.80
	.4	70.00	72.90	77.10	82.00	85.40	90.10
	.6	82.80	86.40	87.20	87.90	92.90	96.30
.3	.2	67.70	71.20	76.10	76.00	82.60	86.40
	.3	68.90	75.10	75.80	80.90	85.10	88.40
	.4	75.10	79.10	80.30	85.60	87.10	90.70
	.6	85.40	87.80	89.10	90.60	94.30	96.10

Table 16

Power to Detect β_{11} When N2 = 25

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	86.30	87.90	91.30	92.80	95.40	97.70
	.3	87.90	89.50	92.00	92.60	95.80	97.90
	.4	90.00	93.30	93.40	95.10	97.00	99.10
	.6	95.10	97.30	97.80	98.70	98.80	99.60
.3	.2	85.40	91.00	91.90	94.40	96.60	97.90
	.3	88.90	92.30	93.20	94.60	96.50	98.20
	.4	90.50	91.70	95.50	96.70	98.10	98.70
	.6	97.10	98.90	98.70	99.30	98.70	99.70

Table 17

Power to Detect β_{11} When N2 = 35

		Level-1 Sample Size					
ρ_{XY}	ρ_{TimeY}	10	15	20	30	50	80
.2	.2	94.90	96.30	97.20	97.90	98.70	100.00
	.3	95.80	97.00	98.50	99.00	99.20	99.80
	.4	97.10	97.00	98.10	99.20	99.50	99.90
	.6	98.60	99.30	99.80	99.80	100.00	100.00
.3	.2	95.00	98.50	97.70	98.00	99.30	99.80
	.3	95.80	98.30	99.20	99.50	99.40	99.90
	.4	96.50	98.70	99.00	99.70	99.50	100.00
	.6	99.40	99.60	99.60	100.00	100.00	100.00

Overall, both predictor-outcome correlations had great impact on the power to detect β_{11}. This was expected as β_{11} is the fixed effect that models the influence of the cross-level interaction on the outcome variable. All conditions with N2 $\geq$ 25 produced at least adequate power to detect β_{11}.

Power to detect the level-2 variances. Changing the correlations between the outcome variable and either predictor had no discernable effect on the power of detecting τ_{00} or τ_{11}. To illustrate this, the results for the combination of the lowest and highest correlations are presented. Tables 18 and 19 show the pattern of the power to detect τ_{00} for the lowest and highest predictor-outcome correlation combinations tested. Tables 20 and 21 illustrate the pattern for τ_{11}.

For the conditions tested, the power to detect τ_{00} depended upon the level-1 and level-2 sample size. Power was 0 for all conditions where N2 was 5. Doubling N2 to 10 did not improve the situation; power was still less than 1% even for large level-1 sample sizes. Increasing N2 beyond 10 produced rapid increases in power, with the majority of conditions reaching adequate, high, or perfect power. Increasing the level-1 sample size to at least 30 did allow for adequate or high power for N2 = 15. When N2 = 25, 15 measurements were sufficient for high power. For the lowest level-1 sample size of 10, adequate power was reached for the two highest level-2 sample sizes tested. In addition, no conditions tested in which N1 = 10 reached power rates above 90%. Overall, high power was achieved in more than 25% of the conditions; perfect power was reached for about a quarter of the tested conditions. In general, power increased faster with an increase in level-2 sample size than with an increase in level-1 sample size (Table 18).

Table 18

Power to Detect τ_{00} When ρ_{TimeY} = 0.2 and ρ_{XY} = 0.2

	N1					
N2	10	15	20	30	50	80
5	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.10	0.50
15	18.90	40.50	61.40	85.40	96.90	99.70
20	48.00	79.10	92.80	99.00	99.90	100.00
25	70.70	91.40	98.10	99.80	100.00	100.00
30	82.90	97.20	99.50	100.00	100.00	100.00
35	88.40	99.60	100.00	100.00	100.00	100.00

Table 19

Power to Detect τ_{00} When ρ_{TimeY} = 0.6 and ρ_{XY} = 0.3

	N1					
N2	10	15	20	30	50	80
5	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.10	0.40
15	19.90	41.40	61.80	86.00	97.00	99.80
20	49.90	79.10	91.20	99.20	100.00	100.00
25	69.20	93.20	98.00	100.00	100.00	100.00
30	80.20	96.40	99.50	100.00	100.00	100.00
35	89.10	98.90	100.00	100.00	100.00	100.00

Power to detect τ_{11} was very similar to detecting power for τ_{00}. The only condition which produced adequate power for τ_{11} but not τ_{00} was the sample size combination of (15, 20). Also, perfect power for τ_{11} was reached for fewer conditions with a level-1 sample size of 30.

Table 20

Power to Detect τ_{11} When $\rho_{TimeY} = 0.2$ and $\rho_{XY} = 0.2$

	N1					
N2	10	15	20	30	50	80
5	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.30
15	19.40	42.10	60.90	86.90	97.70	99.50
20	49.60	80.40	91.80	99.10	100.00	100.00
25	71.90	93.90	98.10	99.80	100.00	100.00
30	82.80	97.80	99.30	99.90	100.00	100.00
35	91.10	99.10	100.00	100.00	100.00	100.00

Table 21

Power to Detect τ_{11} When $\rho_{TimeY} = 0.6$ and $\rho_{XY} = 0.3$

	N1					
N2	10	15	20	30	50	80
5	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.60
15	18.90	43.00	64.80	85.50	97.80	99.00
20	50.40	81.00	91.20	98.20	100.00	100.00
25	70.00	93.20	98.20	99.80	100.00	100.00
30	83.40	97.90	99.70	100.00	100.00	100.00
35	89.90	99.40	99.70	100.00	100.00	100.00

Power to detect the level-1 variance. Under all conditions tested with N2 ≥ 10, power to detect the level-1 variance was perfect, independent of the predictor-outcome correlations and sample sizes. When the level-2 sample size was set to 5, some of the conditions produced power slightly less than perfect. This occurred when the level-1 sample size was either 10 or 15. Under these conditions, power rates were above 99%.

Type I Error Rates. For τ_{01}, all Type I error rates were below 5%, irrespective of correlations and sample sizes. Increasing the level-1 sample size had no discernable effect on the Type I error rates for τ_{01}. For all conditions with N2 being 10, the Type I error rate was 0.10% or 0%. The error rates increased steadily with the increase of N2 for all tested conditions. The pattern does not indicate whether increasing N2 beyond the chosen maximum could produce undesirably high Type I error rates for τ_{01}.

The Type I error rate of the autocorrelation factor was around the expected 5% (M = 5.08, SD = 0.706), but slightly inflated. It was not influenced by any of the variables that were varied in this study. This means that neither the sample size combination nor the strengths of the predictor-outcome correlations had any impact on the Type I error rate of the autocorrelation factor.

Conditions With Nonzero Autocorrelation

Biases. For the conditions tested where N2 was greater than 5, the biases of the fixed effects were all below 5% regardless of the magnitude of the autocorrelation factor. For those conditions where N2 was 5, most biases were below 5%. Some values exceeded 5% but all were below 10%. No discernable pattern that could explain the marginally inflated biases was present.

Conditions tested with nonzero autocorrelation showed inflated estimates for the level-2 error variances τ_{00} and τ_{11} when the total sample size was small (Table 15). The inflation of the biases increased further for conditions with non-zero autocorrelation. Furthermore, the higher the autocorrelation, the higher was the inflation of the estimates of the level-2 variances. Therefore, more conditions produced biases of τ_{00} and τ_{11} beyond 5% and 10%. Under the highest tested autocorrelation, all biases of τ_{00} were greater than 10% for a level-1 sample size of 10. The average inflation of the intercept variance τ_{00} for the lowest sample size combination and the highest autocorrelation factor tested was above 75%.

For all conditions tested in which N2 = 5, the biases were inflated. Only large level-1 sample sizes produced biases in the estimates of τ_{00} that were around or below 5%. Biases were below 5% in less than 7% of the conditions. This was not achieved for the conditions with $\rho^2_{auto} \geq .20$ (Table 22).

Table 22

Biases for τ_{00} When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 5

	N1					
ρ^2_{auto}	10	15	20	30	50	80
0	25.844	15.370	9.314	8.633	5.159	3.640
.05	47.755	32.310	21.220	13.293	6.617	5.247
.10	66.169	43.448	30.904	11.870	7.859	5.336
.15	76.404	50.191	35.433	27.589	12.462	4.036
.20	99.842	60.467	54.616	30.933	19.731	8.956
.25	114.653	80.138	57.789	39.523	17.803	12.000

Increasing the level-2 sample size to 10 produced more than 63% of the tested conditions with desirable levels of biases. The conditions with the highest ρ^2_{auto} needed 50 or 80 measurements per person to decrease the biases below 5%. No conditions with (10, 10) produced high estimates of the intercept variance (Table 23).

Table 23

Biases for τ_{00} When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 10

	N1					
ρ^2_{auto}	10	15	20	30	50	80
0	8.4221	1.651	1.097	1.167	0.959	-0.808
.05	20.441	4.384	0.835	1.153	0.440	1.053
.10	24.220	13.115	2.336	1.900	-0.161	7.197
.15	38.101	13.791	8.534	6.077	1.184	-2.390
.20	52.315	22.454	12.528	2.828	4.783	-0.970
.25	75.270	29.266	17.738	10.608	0.431	2.968

For even higher level-2 sample sizes, the impact of the autocorrelation factor on the intercept variance estimates lessened (Table 24). The effect decreased further for N2 = 25 to the point where only conditions with high levels of autocorrelation and a level-1 sample size of 10 produced undesirably high biases (Table 25).

Table 24

Biases for τ_{00} When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 15

	N1					
ρ^2_{auto}	10	15	20	30	50	80
0	0.405	-0.713	-2.884	-0.164	-0.954	0.134
.05	7.913	1.777	1.388	0.731	-0.005	-0.502
.10	13.699	4.604	2.604	2.210	-1.070	-0.158
.15	17.727	9.525	-0.113	0.996	0.336	-0.318
.20	28.868	12.998	3.179	1.157	-0.359	0.198
.25	42.742	15.116	6.565	4.346	1.477	-1.020

Table 25

Biases for τ_{00} When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 25

	N1					
ρ^2_{auto}	10	15	20	30	50	80
0	0.820	0.001	-0.975	-0.362	-1.534	-0.281
.05	4.930	-0.570	-1.382	1.285	0.663	0.081
.10	4.188	0.597	-2.657	-0.354	0.017	0.365
.15	7.937	-0.462	-2.346	-0.789	-0.870	-1.618
.20	13.691	2.467	-2.126	-0.299	-0.865	-0.377
.25	21.828	1.495	-0.594	-0.299	-0.878	1.3044

For larger level-2 sample sizes, the pattern continued. When N2 = 30, two values exceeded 5%. For N2 = 35, only the largest tested autocorrelation in combination with

the lowest level-1 sample size produced an undesirably large bias in the intercept variance.

The biases for the slope variance τ_{11} behaved similarly to those of the intercept variance. The higher the autocorrelation was, the greater was the inflation of the estimates. Increasing the level-2 sample size produced more conditions with desirable biases below 5%. For identical conditions the biases for τ_{11} were lower than those for τ_{00} (Tables 26 – 28). For all conditions in which the level-2 sample size was 25 or larger, all biases for the intercept variance were below 5%.

Table 26

Biases for τ_{11} When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 5

	N1					
ρ^2_{auto}	10	15	20	30	50	80
0	32.8617	18.8305	12.5967	7.6518	4.8994	2.919
.05	47.852	25.442	16.899	11.440	9.4120	2.841
.10	51.0686	35.2764	21.6302	10.1906	10.052	5.7171
.15	65.3546	40.416	33.9444	24.789	12.4155	6.4905
.20	58.2468	47.2195	38.1037	25.3437	19.6062	9.0961
.25	75.0353	59.6763	46.7008	29.5759	18.5009	8.1358

Table 27

Biases for τ_{11} When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 10

	N1					
ρ^2_{auto}	10	15	20	30	50	80
0	4.042	-1.385	-2.783	0.623	-0.086	2.106
.05	11.506	3.655	4.175	1.499	1.417	-0.372
.10	13.529	7.981	5.002	-0.636	0.063	1.173
.15	18.554	10.299	6.566	-0.551	-1.193	-1.187
.20	29.229	9.989	6.711	5.186	-0.380	-0.890
.25	31.278	14.395	7.346	5.224	0.715	0.567

Table 28

Biases for τ_{11} When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 15

	N1					
ρ^2_{auto}	10	15	20	30	50	80
0	0.237	1.434	-0.599	1.990	-0.029	-2.019
.05	5.9213	1.039	0.767	1.007	-1.412	-0.699
.10	6.772	2.119	2.411	-3.284	0.217	-0.850
.15	8.405	1.149	0.809	1.509	-1.401	-0.193
.20	10.22	1.561	-0.858	-1.142	0.023	-0.064
.25	11.673	7.268	0.9041	-1.468	-2.310	-0.702

While the biases for the level-1 error variance were all below 5% when the autocorrelation factor was set to 0, some conditions with non-zero autocorrelation

produced large negative biases. The estimates were severely deflated for some conditions with low sample size combinations and high autocorrelation.

For a level-2 sample size of 5, about half of the estimates were deflated by more than 5%. The majority of the biases above 5% were higher than 10%, with the highest exceeding 30%. As Table 29 illustrates, even the biases below 5% showed a pattern of becoming more negative with increasing autocorrelation.

Table 29

Biases of σ_e^2 When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 5

	N1					
ρ^2_{auto}	10	15	20	30	50	80
.05	-8.719	-2.998	-1.684	-0.809	-0.453	0.004
.10	-14.489	-6.400	-2.690	-1.371	-0.898	0.047
.15	-19.207	-8.219	-5.398	-2.784	-1.033	-0.033
.20	-25.34	-12.764	-7.951	-3.636	-2.243	-0.805
.25	-31.086	-17.546	-10.819	-5.426	-1.676	-1.193

Similar to the biases of the level-2 variances, increasing the level-2 sample size improved the estimates of the level-1 variance. At N2 = 10, fewer conditions produced undesirable biases, with only two conditions having biases in excess of 10% (Table 30) with one being larger than 20%. This trend of improved biases continued for N2 being 15. Only the condition with the highest autocorrelation and the lowest level-1 sample size had a negative bias of σ_e^2 larger than 10% (Table 31).

Table 30

Biases of σ_e^2 When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 10

	N1					
ρ_{auto}^2	10	15	20	30	50	80
.05	-3.482	-1.425	-0.567	-0.061	-0.075	0.035
.10	-5.792	-1.954	-1.316	-0.196	-0.109	0.190
.15	-8.712	-3.800	-1.802	-1.062	-0.746	-0.278
.20	-15.191	-5.699	-3.809	-0.944	-0.806	-0.188
.25	-20.077	-9.114	-5.845	-1.897	-1.101	-1.216

Table 31

Biases of σ_e^2 When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 15

	N1					
ρ_{auto}^2	10	15	20	30	50	80
.05	-1.788	0.022	-0.221	-0.422	-0.379	0.270
.10	-3.888	-1.055	-0.509	-0.570	-0.301	-0.188
.15	-6.080	-2.851	-1.294	-0.886	-0.311	0.127
.20	-9.766	-4.085	-2.018	-0.515	-0.810	-0.102
.25	-14.064	-6.908	-3.475	-1.839	-0.542	-0.486

For level-2 sample sizes of 15, the biases were further reduced. All of the conditions tested in which N2 ≥10 had biases below 5% when N1 ≥ 20. For the largest level-2 sample size of 35, the largest autocorrelation factor tested still produced a bias of σ_e^2 larger than 10% for a level-1 sample size of 10.

The biases for the autocorrelation factor were mostly below 5%. Similar to the other variances, however, conditions with low sample size combinations and high autocorrelation factors showed undesirably large negative biases in the autocorrelation factor estimates.

For a level-2 sample size of 5, all conditions with level-2 sample sizes of 20 or less produced biases above or around 10%. The estimates for (10, 5) and $\rho^2_{auto} = .25$ were deflated on average by more than 35%. Conditions tested with $\rho^2_{auto} = .25$ produced a bias larger than 5% for N1 = 30 (Table 32). The impact of the magnitude of autocorrelation factor on the estimate of itself was reduced when the level-2 sample size was increased. With the conditions in which N2 = 10, all conditions with a level-1 sample size of at least 20 produced biases below or around 5%. Those conditions in which (10, 10) produced estimates of ρ^2_{auto} deflated between 15% and 21%. (Table 33).

Table 32

Biases of ρ^2_{auto} When $\rho_{TimeY} = .2$ and $\rho_{XY} = .2$ and N2 = 5

	N1					
ρ^2_{auto}	10	15	20	30	50	80
.05	-41.483	-17.317	-9.238	-3.734	-1.564	-0.537
.10	-36.988	-17.456	-9.527	-5.195	-1.952	-0.192
.15	-38.812	-17.058	-9.505	-4.620	-2.488	-0.959
.20	-36.604	-19.613	-11.011	-4.943	-3.484	-1.183
.25	-37.629	-19.440	-11.862	-6.060	-2.212	-1.502

Table 33

Biases of ρ^2_{auto} *When* $\rho_{TimeY} = .2$ *and* $\rho_{XY} = .2$ *and N2 = 10*

	N1					
ρ^2_{auto}	10	15	20	30	50	80
.05	-18.120	-6.781	-2.690	-1.242	-0.381	-1.378
.10	-16.240	-5.715	-2.422	-0.863	-0.524	0.139
.15	-17.100	-7.690	-3.480	-2.372	-1.206	-0.358
.20	-20.560	-7.765	-4.671	-1.753	-1.072	-0.014
.25	-20.620	-9.554	-5.975	-2.368	-1.162	-0.974

Table 34

Biases of ρ^2_{auto} *When* $\rho_{TimeY} = .2$ *and* $\rho_{XY} = .2$ *and N2 = 15*

	N1					
ρ^2_{auto}	10	15	20	30	50	80
.05	-8.647	-3.985	-1.930	-0.489	0.144	0.295
.10	-11.160	-4.441	-1.262	-0.762	-1.112	-0.653
.15	-11.560	-4.608	-2.079	-1.368	-0.391	-0.573
.20	-13.310	-5.305	-2.873	-1.087	-0.947	-0.410
.25	-14.050	-6.361	-3.526	-1.966	-0.368	-0.639

At a N2 = 15, most conditions where N1 ≥ 15 were deflated by less than 5%. Some were deflated slightly more than 5% but less than 7% (Table 34). For larger level-2 sample sizes, only conditions with N1 of 10 produced undesirably high biases in the

autocorrelation factor estimate. For the largest autocorrelation factor tested, this was still the case for N2 = 35.

Power. Within any correlation combination (ρ_{TimeY}, ρ_{XY}) the level of autocorrelation had no impact on the power of any of the fixed effects. This means that the power to detect β_{00} was perfect for all conditions, independent of sample size, the level of autocorrelation, and the values of the predictor-outcome correlations. The power rates to detect the other fixed effects (β_{10}, β_{01}, and β_{11}) were similar in magnitude and pattern to those presented in the section on the conditions with zero autocorrelation.

The level of autocorrelation had great impact on the power to detect the level-2 variances τ_{00} and τ_{11}. In general, an increase in ρ^2_{auto} resulted in a lower power of both level-2 variances, holding all other variables constant. This resulted in many more conditions where power for τ_{00} and τ_{11} did not reach acceptable values. Conditions in which $\rho^2_{auto} \geq .1$ produced unacceptable power values for a level-2 sample size of 15 (Table 35), when the level-1 sample size was below 80. Specifically, less than 20% of the conditions in which N2 = 15 produced adequate or good power.

Table 35

Power to Detect τ_{00} When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 15

	N1					
ρ^2_{auto}	10	15	20	30	50	80
0	18.90	40.50	61.40	85.40	96.90	99.70
.05	10.20	18.50	32.00	57.10	85.00	97.40
.1	7.40	10.80	19.30	41.00	74.40	94.40
.15	6.80	8.60	10.40	27.30	60.40	86.20
.2	5.70	6.10	7.60	17.30	46.50	77.70
.25	5.60	5.00	5.50	11.70	34.10	65.50

Increasing the level-2 sample size produced fewer undesirable power rates. The level-1 sample sizes, however, also had to be increased to compensate for an increase in the autocorrelation factor (Table 36). This trend continued for the conditions with the highest level-2 sample size tested. When N2 = 35, squared autocorrelation factors of .15 required a level-1 sample size of more than 20 to reach adequate or high power. Almost 50% of conditions tested produced inadequate power rates when N2 was 35 (Table 37).

Table 36

Power to Detect τ_{00} When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 25

	N1					
ρ^2_{auto}	10	15	20	30	50	80
0	70.70	91.40	98.10	99.80	100	100
.05	43.70	68.30	86.30	97.80	99.90	100
.1	31.30	51.10	71.70	93.50	99.70	99.90
.15	23.40	37.30	55.70	84.00	98.80	100
.2	18.60	30.60	43.10	73.10	95.80	99.80
.25	17.60	20.90	33.40	59.40	89.50	99.40

Table 37

Power to Detect τ_{00} When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 35

	N1					
ρ^2_{auto}	10	15	20	30	50	80
0	88.40	99.60	100	100	100	100
.05	62.00	87.90	97.00	99.80	100	100
.1	49.30	76.10	91.20	99.20	100	100
.15	34.30	61.70	78.60	95.30	100	100
.2	28.00	48.50	67.20	92.00	99.40	100
.25	25.90	36.20	53.20	84.10	98.30	99.90

The effect of the magnitude of the autocorrelation factor on the power to detect the level-2 variances was similar in pattern for all the predictor-outcome correlations

tested. There was some variance, but no pattern was detectable. Also, the decrease in power is similar for τ_{00} and τ_{11}.

In the presence of autocorrelation, the power to detect the level-1 variance was reduced in the smaller sample size combinations mostly for N1 = 10 but a few with N1 = 15. Power was not influenced by the predictor-outcome correlations. It was, however, dependent upon the magnitude of the autocorrelation factor as well as the sample sizes. Power decreased with the increase of the autocorrelation factor and the decrease of either level's sample size. The effect produced some power values below 90% (minimum 83.70% for (10, 5), ρ_{TimeY} = .3, and ρ_{XY} = .2,) for the lowest sample size combination and the highest autocorrelation factor (Table 38).

Table 38

Power to Detect σ_e^2 *When* ρ_{TimeY} *= .2 and* ρ_{XY} *= .2 and N1 = 10*

	N2						
ρ^2_{auto}	5	10	15	20	25	30	35
0	100	100	100	100	100	100	100
.05	97.50	100	100	100	100	100	100
.1	96.70	99.70	100	100	100	100	100
.15	91.10	99.50	99.90	100	100	100	100
.2	89.40	98.70	99.40	100	100	100	100
.25	85.00	96.80	98.90	99.70	100	100	100

In the conditions tested, the strengths of the predictor-outcome correlations did not impact the power to detect the autocorrelation. The variables that changed the power to detect the autocorrelation factor were the magnitude of the autocorrelation factor itself

and the sample sizes. The power to detect the autocorrelation factor increased with the increase in the autocorrelation factor. Similarly, larger sample sizes on either level increased the power to detect $\rho_{auto.}$ The following discussion uses the lowest predictor-outcome correlation combination, that is $\rho_{TimeY} = .2$ and $\rho_{XY} = .2$, as a representative example. The tables for the other predictor-outcome correlation combinations contain similar values with some nonsystematic fluctuation.

Power to detect the autocorrelation factor was high for about half of the conditions in which N2 = 5. Increasing the number of measurements to 50 guaranteed high power to detect the autocorrelation factor. Even for N2 = 5, several conditions produced perfect power (Table 39).

Table 39

Power to Detect ρ^2_{auto} *When* $\rho_{TimeY} = .2$ *and* $\rho_{XY} = .2$ *and N2 = 5*

	N1					
ρ^2_{auto}	10	15	20	40	50	80
.05	4.60	20.50	40.50	68.00	90.90	99.30
.1	9.30	41.40	68.10	91.80	99.60	100
.15	12.40	58.00	84.20	98.50	100	100
.2	19.90	70.40	92.00	99.50	100	100
.25	22.20	78.60	95.90	100	100	100

Increasing the level-2 sample size to 10 produced adequate or high power results for those conditions where N1 was 15 or greater and the square of the autocorrelation factor was .10 or higher. A further increase in level-2 sample size produced more conditions with high or perfect power. When N2 was set to 20, all but the lowest

combination of N1 and ρ^2_{auto} produced at least adequate power (Table 42). Further increases in level-2 sample size improved the power rates to the point that only the lowest combination of N1 and ρ^2_{auto} had inadequate power, and more than two-thirds of the conditions had perfect power. For conditions where N2 was 35, even the lowest N1 and ρ^2_{auto} produced adequate power of 81.90%.

Table 40

Power to Detect ρ^2_{auto} When ρ_{TimeY} = .2 and ρ_{XY} = .2 and N2 = 10

	N1					
ρ^2_{auto}	10	15	20	40	50	80
.05	19.60	51.30	74.70	93.40	99.70	100
.1	39.20	83.70	95.40	99.80	100	100
.15	55.10	94.00	99.40	100	100	100
.2	63.10	98.00	100	100	100	100
.25	74.40	99.60	100	100	100	100

Table 41

Power to Detect ρ^2_{auto} *When* ρ_{TimeY} *= .2 and* ρ_{XY} *= .2 and N2 = 20*

	N1					
ρ^2_{auto}	10	15	20	40	50	80
.05	53.80	85.90	96.30	99.80	100	100
.1	81.00	99.10	100	100	100	100
.15	91.70	99.90	100	100	100	100
.2	97.70	100	100	100	100	100
.25	99.00	100	100	100	100	100

Table 42

Power to Detect ρ_{auto} *When* ρ_{TimeY} *= .2 and* ρ_{XY} *= .2 and N2 = 30*

	N1					
ρ^2_{auto}	10	15	20	40	50	80
.05	74.10	96.60	99.70	100	100	100
.1	94.60	99.80	100	100	100	100
.15	98.90	100	100	100	100	100
.2	99.80	100	100	100	100	100
.25	100	100	100	100	100	100

Type I errors. The only variable tested for Type I error under non-zero autocorrelations was τ_{01}. The power values were similar to those found under zero autocorrelation conditions. None exceeded 5%. In general, the power rates were lower for lower level-2 sample sizes.

Chapter 5

Discussion

The primary focus for this study was on the effects of level-1 and level-2 sample sizes on both the accuracy of parameter estimates and the statistical power in hierarchical linear growth modeling when more measurements are taken from fewer people. Overall, the results of this study do not confirm those of other studies that suggest that level-2 sample sizes of 100 or more are necessary to achieve adequate or high power (Hedeker, et al., 1999; Raudenbush & Liu, 2001; Zhang & Wang, 2009). The findings of this study were more in line with Gagné and Estes (2009a, 2009b), who investigated similar conditions in two-level cross-sectional hierarchical linear models, and Ferron et al. (2009), who focused on a very simple AB model. The results overlap somewhat with Jenson, Clark, Kircher, and Kristjansson (2007) who found that high levels of autocorrelation have a major impact on power.

Zero Autocorrelation

Conditions with zero autocorrelation allow for good estimates of the fixed effects and the level-1 variance for sample sizes as low as (10, 5). For small sample size combinations, however, researchers should expect the level-2 variances to be inflated by 30% or more. In general, the biases are independent of the strengths of the predictor-outcome correlations.

Power to detect the fixed effects depends on the strengths of the predictor-outcome correlations as well as the sample sizes for all conditions tested. The stronger the correlations and the larger the sample sizes are, the higher the power to detect the effect. The correlation associated with a fixed effect has a larger impact than those that

are not directly associated with them. The strength of the other predictor outcome correlation, for example ρ_{TimeY} in the case of β_{01}, influences the power to predict the fixed effect somewhat.

In studies where the correlations between Time and Y or X and Y are expected to be small, sample sizes need to be relatively large to achieve adequate power. Specifically, when ρ_{TimeY} and ρ_{XY} are both low, high power can be achieved with 80 measurements from 35 participants. Furthermore, data from 80 measurements from 25 people or at least 30 measurements from 35 participants can have adequate power.

The necessary sample sizes decreases with an increase in predictor-outcome correlations. When both correlations are medium in strength (i.e., .3), the necessary sample sizes to achieve adequate or high power decrease. Adequate power can be reached by taking 30 measurements of 15 people. When 25 people are included in a study, as few as 10 measurements is sufficient to achieve adequate power.

Under conditions in which the correlation between Time and Y is assumed to be high (i.e., around.6) and the correlation between a person-level variable and Y are medium in strength, sample size recommendations are even more lenient. Having 5 participants, adequate or even high power to detect β_{10} can be achieved with 15 or more measurements per person. A study with only 10 participants can produce adequate power to detect all fixed effects when at least 50 measurements are taken from each individual. Involving 15 participants produces adequate power for even 10 measurements per person.

The power to detect the level-2 variances is independent of the predictor-outcome correlations. It depends strongly, however, on the sample sizes. Studies that involve 10 participants or less have no or very little power to detect τ_{00} or τ_{11}. For studies with more

than 10 but less than 30 participants, increasing the number of measurements allows for adequate power to detect the level-2 variances. When at least 30 participants are involved, 10 measurements per person suffice to yield adequate power.

Nonzero Autocorrelation

As autocorrelation is an established complication in studies where many measurements of the same people are collected over relatively short periods of time (Yaffee & McGee, 2000), it needs to be considered when making decisions about sample sizes for single-case designs. This study shows that the magnitude of the autocorrelation factor does not affect the biases of the fixed effects. Increasing the autocorrelation does, however, inflate the level-2 variances. In addition, larger autocorrelation results in the level-1 variances being negatively biased, as well as downwardly biased estimates of the autocorrelation factor itself. In general larger sample sizes produce better estimates of the variances and of the autocorrelation. Increasing the level-2 sample size has a greater impact on the accuracy of the estimates than using larger level-1 samples. Specifically, the inflation of any of the variances was counteracted entirely with level-2 sample sizes of 20 or greater for even the highest magnitude tested.

Similarly to the bias in the estimates of the fixed effects, the autocorrelation factor did not influence the power to detect the fixed effects. Thus, to the extent that statistical power will motivate a researcher's decision, the same sample size recommendations that apply to situations where there is no autocorrelation hold for conditions with non-zero autocorrelation.

The power to detect the level-2 variances, however, is influenced greatly by the degree of autocorrelation. For a level-2 sample size of 35, moderate to high

autocorrelation values are associated with inadequate power values for 20 measurements or less per person. For a level-2 sample size of 15 or less, 50 or more measurements are necessary to achieve adequate or high power.

The influence of the number of measurements on the power to detect the autocorrelation factor was high as well. Higher values of autocorrelation needed at least 20 measurements per person to be detected with at least adequate power when N2 was only 5. Power is adequate or high for conditions with low levels of autocorrelation if the number of measurements is at least 20 for small level-2 sample sizes.

Recommendations for Single-Case Researchers

Overall, under conditions in which the predictor-outcome correlations are moderate or high, the single-case researcher should recruit at least 15 participants and take 80 or measurements from each. If 20 participants can be included in a study, the number of measurements can be decreased to 50. Under these conditions, the researcher can expect unbiased estimates of all parameters and adequate or high power to detect all fixed and random effects in the presence of low to moderate autocorrelation factors ($\rho^2_{auto} \leq .20$). If very high levels of autocorrelation ($\rho^2_{auto} = .25$) are expected, at least 80 measurements of 20 participants or 50 measurements of 25 participants should be collected.

Furthermore, the results provide a clear recommendation against the use of only 5 participants with only 10 measures per person when using HLM to estimate effects. The magnitudes of the biases in the variances were too large and the statistical power was too low to allow such an analysis to produce meaningful results.

The biases of the fixed effects are reasonably low for studies including 5 participants or more. HLM can thus be a useful tool in establishing the values of the fixed effects for as few as 5 participants. The single-case researcher is encouraged to take as many measurements per person as possible in order to increase power. The estimates of the variances need to be considered with caution when the sample sizes are low. Still, using HLM on a small sample of participants can indicate whether a larger scale study is warranted. Using larger numbers of participants and/or numbers of measurements can yield meaningful results, depending on the combination of outcome-predictor correlations and autocorrelation. Fifteen participants can be a reasonable level-2 sample size when the predictor-outcome correlations are high.

Future research

The present study used values for the level-1 sample size that are common in single-case research. Future methodological research should investigate minimum sample size combinations that are more typical in group research, that is, $N1 \leq 10$ and $N2 > 35$. Using only small or medium correlations between Time and Y would also be a useful way to obtain results more relevant to group-design research.

The parameters for this study should also be expanded to more conditions that are applicable in single-case research. As discussed previously, power rates for the fixed effects as well as for the random effects increased substantially with the jump from N2 = 10 to N2 = 15. Thus, future studies should increase the level-2 sample size in this range in small increments of 2 or even 1 to investigate the patterns more precisely. In addition, the results of this study indicate that small sample size combinations yield adequate or high power rates if the correlations between the predictor and the outcome variable are

sufficiently large. Increasing the correlation between the level-2 predictor and Y beyond the scope of this study may yield valuable insights into power.

A limitation of this study was that the correlation between Time and X and the correlation between the cross-level interaction and Y were both fixed, the former at 0, the latter at .3. Future research should include several values of the correlation between the cross-level interaction and Y to test the impact of multicollinearity on the biases and power of the parameters. It should also consider a variety of magnitudes of the correlation between Time and X to gain insight into explaining and predicting between person variance in growth.

To get the most out of the application of HLM, larger sample sizes are preferred. In many situations, however, sample sizes of 15 or more are unrealistic for single-case researchers. This does not exclude the use of HLM. Several similar studies can be combined through HLM meta-analyses (Morgan & Sideridis, 2006; Raudenbush & Bryk, 2002; Van den Noortgate & Onghena, 2003b). Future studies should look at three-level models that suit single-case meta-analyses and investigate minimum sample size combinations.

The results of this study also pose some questions for methodological researchers. The Type I error rate of τ_{01} increased systematically without reaching 5% when N2 was increased. Investigations into the behavior of τ_{01} for level-2 sample sizes larger than 35 would be of interest to determine whether the error rates become undesirably large. Furthermore, the non-perfect power for the level-1 error variance is a curious result. While no condition with zero autocorrelation produced power rates of σ_e^2 below 99%, in this study it was found that low sample sizes can produce situations where the level-1

error variance might be reported to be not statistically significantly different than 0. The large impact on the power of σ_e^2 in the presence of autocorrelation, producing some power rates below 90%, complicates these findings and warrants future investigation as well.

References

Alberto, P., & Troutman, A. C. (2009). *Applied behavior analysis for teachers* (8th ed.). Upper Saddle River, NJ: Merrill/Pearson.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case*. *Behaviour Research and Therapy, 31*, 621-631.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change*. New York, NY: Pergamon Press.

Bengali, M. K., & Ottenbacher, K. J. (1998). The effect of autocorrelation on the results of visually analyzing data from single-subject designs. *The American Journal of Occupational Therapy, 52*, 650-655.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics, 15*, 291-420.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10*, 229-242.

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1966). *Experimental and quasi-experimental designs for research*. Chicago, IL: McNally.

Center, B. A., Skiba, R. J., & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design reserach. *Journal of Special Education, 19*, 387-400.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *41*, 372-384.

Ferron, J. M., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research, 37*, 379-403.

Franklin, R. D., Gorman, B. S., Beasley, M. T., & Allison, D. B. (1997). Graphical display and visual analysis. In R. D. Franklin, D. B. Allison & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119-158). Mahwah, NJ: L. Erlbaum Associates.

Gagné, P., & Estes, K. E. (2009a). *The effect of predictor-criterion correlation magnitude on sample size recommendations in HLM*. Paper presented at the American Educational Research Association, San Diego.

Gagné, P., & Estes, K. E. (2009b). *The variance of the regression R^2 across level-2 units in HLM*. Paper presented at the American Educational Research Association, San Diego.

Haardörfer, R., & Gagné, P. (in press). The use of randomization tests in single-subject research. *Focus on Autism and Other Developmental Disabilities*.

Hedeker, D., Gibbons, R. D., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics, 24*, 70-93.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research, 26*, 329-367.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7* (107-118).

Huitema, B. E. (1988). Autocorrelation: 10 years of confusion. *Behavioral Assessment, 10*, 253-294.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools, 44*, 483-493.

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of applied behavior analysis, 11*, 277-283.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

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

Kwok, O., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research, 42*, 557-592.

Lumpkin, P. W., Silverman, W. K., Weems, C. F., Markham, M. R., & Kurtines, W. M. (2002). Treating a heterogeneous set of anxiety disorders in youths with group

cognitive behavioral therapy: A partially nonconcurrent multiple-baseline evaluation. *Behavior Therapy, 33*, 163-177.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 86-92.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341-351.

Mok, M. (1995). Sample size requirements for 2-level desgins in educational research. *Multilevel Modelling Newsletter, 7*(2), 11-15.

Morgan, P. L., & Sideridis, G. D. (2006). Contrasting the effectiveness of fluency interventions for students with or at risk for learning disabilities: A multilevel random coefficient modeling meta-analysis. *Learning Disabilities Research & Practice, 21*, 191-210.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425. (2002).

O'Connell, A. A., & McCoach, D. B. (2008). *Multilevel modeling of educational data*. Charlotte, NC: IAP.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173-185.

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

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*, 199-213.

Raudenbush, S. W., & Liu, X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*, 387-401.

SAS Institute. (2004). *SAS/IML 9.1 user's guide*: SAS Publishing.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation, 113*, 95-109.

Sharpley, C. F., & Alavosius, M. P. (1988). Autocorrelation in behavioral data: An alternative perspective. *Behavioral Assessment, 10*, 243-251.

Sideridis, G. D., & Greenwood, C. R. (1997). Is human behavior autocorrelated? An empirical analysis. *Journal of Behavioral Education, 7*, 273-293.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*, 237-259.

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

Suen, H. K. (1987). On the epistemology of autocorrelation in applied behavior analysis. *Behavioral Assessment, 9*, 113-124.

Suen, H. K., & Ary, D. (1987). Autocorrelation in applied behavior analysis: Myth or reality? *Behavioral Assessment, 9*, 125-130.

Todman, J. B., & Dugard, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests*. Mahwah, NJ: Erlbaum.

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*, 325-346.

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35*, 1-10.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*, 1-15.

Yaffee, R., & McGee, M. (2000). *Introduction to time series analysis and forecasting with applications of SAS and SPSS*. New York, NY: Academic Press.

Zhang, Z., & Wang, L. (2009). Statistical power analysis for growth curve models using SAS. *Behavior Research Methods, 41*, 1083-1094.

Zucker, D. R., Schmid, C. H., McIntosh, M. W., D'Agostino, R. B., Selker, H. P., & Lau, J. (1997). Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *Journal of Clinical Epidemiology, 50*, 401-410.