INVESTIGATING INTERVENTION EFFECT OF A READING PROGRAM FOR LOW-ACHIEVING INCARCERATED YOUTH INCLUDING SIMULATION STUDIES FOR LONGITUDINAL RESEARCH

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Jing Zhu, M.A., M.A.S.

*****

The Ohio State University
2008

Dissertation Committee:

Professor Ayres G. D’Costa, Adviser
Professor Dorinda J. Gallant
Professor Richard G. Lomax
Professor William E. Loadman

Approved by

-------------------------------------------

Adviser
College of Education and Human Ecology
ABSTRACT

The majority of incarcerated youth are unsuccessful in school and many have a significant reading deficit. This study aimed to determine if the Scholastic READ 180 program had a meaningful impact on the reading proficiency of low-achieving incarcerated youth in a large mid-western state, when salient subject covariates were controlled for their influences. The study was based on a longitudinal experimental design in which the eligible youth were randomly assigned to either the READ 180 program or a comparison group being instructed by a traditional reading program. The course of investigation lasted for one school year during which the subjects were measured for their reading proficiency by the Scholastic Reading Inventory (SRI) prior to the treatment and again at the end of each of the four terms. The mixed-effects models were applied to the sample data with a maximum of five repeated measures. Results indicated that subjects exposed to the READ 180 program demonstrated accelerated reading growth over time, although it took time for them to display marked reading gain. In addition, it was found that the variability in the initial reading status of the low-performing incarcerated youth could be attributed to covariates including age, disability status, and another baseline reading test, while a baseline math assessment could account for the variability in a constant growth rate
for reading proficiency over time. It also seemed that at higher grade levels, the reading growth of these youth was expected to decelerate over time.

Another major focus of the study concerned identifying statistical properties of the probability distribution of the estimate for the READ 180 intervention effect on low-performing incarcerated youth as well as methodologically guiding future longitudinal research with power analysis. Based upon both the bootstrap and the Monte Carlo approaches, it was believed that the intervention effect estimate was modeled quite well using the original sample with sufficient power. Power simulations via the Monte Carlo method indicated that statistical power of detecting true treatment effect is substantially influenced by the magnitude of the effect and the sample size but not the number of points in time, given that highly unbalanced data pattern is expected for similar longitudinal studies.

The present study both contributed to empirical studies of the READ 180 program impact on low-achieving incarcerated youth, and to methodological considerations for modeling the intervention effect. Implications of the study and recommendations for future research and policy were also discussed.
Dedicated

to my husband Xu, Hua

and

to my parents Zhu, Jianguo and Yang, Dongmei
ACKNOWLEDGMENTS

I wish to express sincerest thanks to my adviser, Dr. Ayres D’Costa, for his continuous guidance and encouragement throughout the past years. Without his invaluable support at the final stage of my doctoral study, I would not have been able to complete this dissertation as well as the program.

I am greatly indebted to Dr. William Loadman for important discussions of my dissertation project. In particular, I wish to thank him for being considerate and supportive when I started a new dissertation topic. His efforts in addressing my questions and providing useful resources in time made it easier to finish this study.

My grateful appreciation is also extended to Dr. Richard Lomax for his outstanding scholarship, generous support, and patience. His time and help means a lot to me. I am thankful for the insights shared by Dr. Lomax as a consultant for the READ 180 project. I owe many thanks to Dr. Dorinda Gallant for serving on my dissertation committee and her efforts in arranging a videoconference for my defense when she is out of town on vacation.

My sincere thanks go to Dr. Robert Cudeck for the stimulating advice on longitudinal data analysis and all his time and support since I started this study, to Dr. Michael Browne for sitting in on my candidacy examination committee and all his help with finding a summer internship, to Drs. Elizabeth Stasny and Doug Wolfe for
serving as my committee for the M.A.S. degree and also for their help with finding jobs in the field of statistics, and to Dr. Bruce Kimball for his kindness in quickly approving the petition regarding the arrangements for my defense.

I greatly appreciate all the help provided by Deb Zabloudil since I joined the School of Educational Policy and Leadership. I gratefully thank Francisco Gómez-Bellengé, former director of the National Data Evaluation Center, for his friendship and support. My thanks also go to all the other NDEC employees, Susan Ding, Jeff Brymer-Bashore, Michele Patton, and Scott Shanline.

I want to thank my friends Feifei Ye, Meiling Tang, Xueli Wang, Jie Yao, Yan Wang, Raeal Moore, Melinda Rybski, Wenbin Qi and Yun Gong. Thank you for all your help and encouragement!

I especially want to thank my dear husband Hua. Without his unconditional love and persistent care during the past six years, I would not have embarked on and finished this wonderful but challenging journey. I also owe big thanks to my parents, my grandmother, and my sister for their love, support, and faith in me.

Finally, I want to thank all the people who are not explicitly mentioned above but have helped me in various ways.
October 16, 1981.............................Born – Jiangsu, P.R. China

2003........................................B.A., English
Nanjing University, Nanjing, P. R. China

2006.........................................Graduate Interdisciplinary Specialization
in Survey Research, Center for Survey
Research, The Ohio State University

2007.........................................M.A., Quantitative Research,
Evaluation, and Measurement
The Ohio State University

2007.........................................M.A.S., Applied Statistics
The Ohio State University

2003 – 2004................................Graduate Fellow
The Ohio State University

2004 – 2006................................Graduate Teaching Associate
The Ohio State University

2006 – 2008................................Graduate Research Associate
The Ohio State University

FIELDS OF STUDY

Major Field:   Education

Area of Focus:  Quantitative Research, Evaluation, and Measurement
Statistics
Survey Research
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xv</td>
</tr>
<tr>
<td>Chapter 1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Background of the Study</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Purpose of the Research</td>
<td>7</td>
</tr>
<tr>
<td>1.3 Research Questions</td>
<td>9</td>
</tr>
<tr>
<td>1.4 Significance of the Study</td>
<td>10</td>
</tr>
<tr>
<td>1.5 Basic Assumptions</td>
<td>11</td>
</tr>
<tr>
<td>1.6 Organization of the Dissertation</td>
<td>12</td>
</tr>
<tr>
<td>Chapter 2 Review of The Literature</td>
<td>13</td>
</tr>
<tr>
<td>2.1 The Scholastic READ 180 Instructional Model</td>
<td>13</td>
</tr>
<tr>
<td>2.2 Salient Factors regarding Adolescent Reading Achievement</td>
<td>15</td>
</tr>
<tr>
<td>2.3 Statistical Approaches for Longitudinal Data Analysis</td>
<td>17</td>
</tr>
<tr>
<td>2.3.1 Repeated Measures ANOVA</td>
<td>17</td>
</tr>
<tr>
<td>2.3.2 Extended Linear Model for Repeated Measures</td>
<td>20</td>
</tr>
</tbody>
</table>
Chapter 4 Results ........................................................................................................ 91

4.1 Exploratory Analysis ............................................................................................. 93

4.2 Results based on Mixed-Effects Modeling ............................................................ 103

4.2.1 Selecting the Covariance Structure .................................................................. 103

4.2.2 Selecting the Mean Structure .......................................................................... 105

4.2.3 The Selected Final Model with Restricted Maximum Likelihood Estimation ................................................................. 112

4.2.4 Results for the Final Model after Removing the Outliers ............................... 124

4.2.5 Normality Check ............................................................................................. 126

4.3 Bootstrap Results .................................................................................................. 128

4.3.1 Sampling Distribution of the Treatment Effect Estimate Based on the Bootstrap ................................................................................................................................. 128

4.4 Monte Carlo Simulation Results .......................................................................... 132

4.4.1 Comparison with the Bootstrap Analysis ....................................................... 133

4.4.2 Influence of Magnitude of Treatment Effect on Power with Unbalanced Data .................................................................................................................................................................. 138

4.4.3 Influence of Sample Size on Power with Unbalanced Data ......................... 143

4.4.4 Influence of Number of Time Points on Power with Unbalanced Data ................................................................. 148

4.4.5 Power Comparisons for Balanced and Unbalanced Data .............................. 150

4.4.6 Influence of Number of Time Points on Power with Balanced Data .................. 152

Chapter 5 Summary, Discussion, and Conclusion .................................................... 157

5.1 Summary and Discussions .................................................................................... 157
5.2 Implications ........................................................................................................... 167
5.3 Limitations ........................................................................................................... 170
5.4 Recommendations for Future Research: ......................................................... 172
References .................................................................................................................. 174
Appendix A Matlab Script for Bootstrap ................................................................. 189
Appendix B SAS Macro ............................................................................................ 192
Appendix C PERL Script .......................................................................................... 194
Appendix D Matlab Script for Monte Carlo ............................................................. 198
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.1: Demographic Information of All Low-Achieving Incarcerated Youth in a Large Mid-Western State in the School Year 2006-2007.</td>
<td>47</td>
</tr>
<tr>
<td>Table 3.2: Demographic Information of the Low-Achieving Incarcerated Youth Who Received at Least Two Quarters of Treatment in a Large Mid-Western State in the School Year 2006-2007</td>
<td>49</td>
</tr>
<tr>
<td>Table 3.3: Demographic Information of the Low-Achieving Incarcerated Youth Who Received Less than Two Quarters of Treatment in a Large Mid-Western State in the School Year 2006-2007</td>
<td>51</td>
</tr>
<tr>
<td>Table 3.4: Commonly Used Covariance Structures for Residuals</td>
<td>63</td>
</tr>
<tr>
<td>Table 3.5: Correlations between Age, Disability, ReadCAT, MathCAT, Grade Level, and Number of Measurements Based on the Information of the Accessible Population (N=724)</td>
<td>73</td>
</tr>
<tr>
<td>Table 3.6: Counts for Grade Levels at Different Ages</td>
<td>74</td>
</tr>
<tr>
<td>Table 3.7: Joint Probability Mass Function for Age and Grade Level</td>
<td>75</td>
</tr>
<tr>
<td>Table 3.8: Empirical Distribution of the Number of Measurements of the Outcome</td>
<td>80</td>
</tr>
<tr>
<td>Table 3.9: Observed Attrition Rate after Receiving x Measurements</td>
<td>80</td>
</tr>
<tr>
<td>Table 4.1: Mean SRI Scores at Different Measurement Occasions for the READ 180 Group, the Comparison Group, and the Overall</td>
<td>95</td>
</tr>
<tr>
<td>Table 4.2: Standard Deviations of SRI Scores at Different Measurement Occasions for the READ 180 Group, the Comparison Group, and the Overall</td>
<td>95</td>
</tr>
<tr>
<td>Table 4.3: Autocorrelations for Five Repeated Measures of the SRI for All Subjects</td>
<td>96</td>
</tr>
</tbody>
</table>
Table 4.4: Number and Percentage of Subjects with Positive and Negative Growth Slopes in the READ 180 Group, the Comparison Group, and the Overall .............................................................. 100

Table 4.5: Fit Indices for Different Error Covariance Structures $\Sigma$ for Quadratic Models with Full Mean Structure of Fixed Effects Using REML Estimation ................................................................. 105

Table 4.6: Fit Indices for Linear and Quadratic Models with Full Mean Structure of Fixed Effects and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation ................................................................. 107

Table 4.7: Estimates of Significant Fixed Effects for the Linear Model with Full Mean Structure and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation ................................................................. 107

Table 4.8: Estimates of Significant Fixed Effects for the Quadratic Model with Full Mean Structure and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation ................................................................. 108

Table 4.9: Estimated Covariance Matrices of the Random Effects for Linear and Quadratic Models with Full Mean Structure of Fixed Effects and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation .......... 108

Table 4.10: Fit Indices for Two Nested Quadratic Models with Full Mean Structure and Reduced Mean Structure of Fixed Effects and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation .......... 111

Table 4.11: Estimates of Significant Fixed Effects for Two Nested Quadratic Models with Full Mean Structure and Reduced Mean Structure of Fixed Effects and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation .......... 111

Table 4.12: Estimated Covariance Matrices of the Random Effects for Two Nested Quadratic Models with Full Mean Structure and Reduced Mean Structure of Fixed Effects and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation .......... 111

Table 4.13: Fit Indices for the Final Model Using REML Estimation .......... 116

Table 4.14: Estimates for the Fixed Effects in the Final Model Using REML Estimation .......... 116

Table 4.15: Estimated Covariance Matrix for the Random Effects in the Final Model Using REML Estimation .......... 116
Table 4.16: Fit Indices for the Final Model with REML Estimation after Removing Outliers ................................................................. 125

Table 4.17: Estimates for the Fixed Effects in the Final Model with REML Estimation after Removing Outliers ............................................. 125

Table 4.18: Estimated Covariance Matrix for the Random Effects in the Final Model with REML Estimation after Removing Outliers .................. 125

Table 4.19: Mean, Standard Error, and Confidence Intervals Produced by the Bootstrap and SAS PROC MIXED Procedure ................................ 130

Table 4.20: Power Comparison between the Bootstrap and the SAS PROC MIXED .................................................................................................. 131

Table 4.21: Mean, Median, Standard Error, and Confidence Intervals Produced by the Bootstrap and the Monte Carlo with Comparable Design Factors ..................................................................................................................... 135

Table 4.22: Power Comparison between the Bootstrap and the Monte Carlo with Comparable Design Factors ....................................................... 136

Table 4.23: The Corresponding Cohen’s Effect Sizes for the Bootstrap Estimated Quantiles of the Treatment Effect Distribution ...................... 141
<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 3.1: A Schematic Diagram of the Bootstrap in One-Sample Conditions</td>
<td>66</td>
</tr>
<tr>
<td>Figure 3.2: Correlation Matrix Plot for Age, Disability, ReadCAT, MathCAT, Grade Level, and Number of Measurements Based on the Information of the Accessible Population (N=724)</td>
<td>73</td>
</tr>
<tr>
<td>Figure 3.3: Joint Distribution of Age and Grade Level</td>
<td>75</td>
</tr>
<tr>
<td>Figure 3.4: A Schematic Diagram of Monte Carlo Simulation for Balanced Data</td>
<td>79</td>
</tr>
<tr>
<td>Figure 3.5: Linear Regression Relationship between Observed Attrition Rate and Number of Measurements</td>
<td>81</td>
</tr>
<tr>
<td>Figure 3.6: Simulated Distribution for Number of Measurements</td>
<td>83</td>
</tr>
<tr>
<td>Figure 3.7: A Schematic Diagram of Monte Carlo Simulation for Unbalanced Data</td>
<td>84</td>
</tr>
<tr>
<td>Figure 4.1: Time Plot of the Mean Responses for the READ 180 Group, the Comparison Group, and the Overall</td>
<td>95</td>
</tr>
<tr>
<td>Figure 4.2: Autocorrelation Scatter-Plot Matrix for Five Repeated Measures of the SRI for All Subjects</td>
<td>96</td>
</tr>
<tr>
<td>Figure 4.3: Spaghetti Plots for the Overall Group in Total (Left Panel), Subjects with Positive Slopes (Middle Panel), and Subjects with Negative Slopes (Right Panel)</td>
<td>98</td>
</tr>
<tr>
<td>Figure 4.4: Spaghetti Plots for the READ 180 Group in Total (Left Panel), Subjects with Positive Slopes (Middle Panel), and Subjects with Negative Slopes (Right Panel)</td>
<td>99</td>
</tr>
<tr>
<td>Figure 4.5: Spaghetti Plots for the Comparison Group in Total (Left Panel), Subjects with Positive Slopes (Middle Panel), and Subjects with Negative Slopes (Right Panel)</td>
<td>99</td>
</tr>
</tbody>
</table>
Figure 4.6: Subject-Specific Coefficients $R^2_i$ of Multiple Determination and the Overall Coefficient $R^2_{meta}$ of Multiple Determination (Dashed Lines) for First-Stage Models which Assume Linear (Left Panel), Quadratic (Middle Panel), and Cubic (Right Panel) Subject-Specific Profiles

Figure 4.7: Time Plot of the Fitted Mean Difference in the SRI Scores between the READ 180 Group and the Comparison Group for the Final Quadratic Reduced Model Using REML Estimation

Figure 4.8: Mean Functions for the READ 180 and the Comparison

Figure 4.9: Mean Functions for the READ 180 and the Comparison, at the .25 and .75 Quantiles of Age

Figure 4.10: Mean Functions for the READ 180 and the Comparison, with Disability and without Disability

Figure 4.11: Mean Functions for the READ 180 and the Comparison, at the .25 and .75 Quantiles of ReadCAT

Figure 4.12: Mean Functions for the READ 180 and the Comparison, at the .25 and .75 Quantiles of MathCAT

Figure 4.13: Mean Functions for the READ 180 and the Comparison, at the .25 and .75 Quantiles of Grade Level

Figure 4.14: Trellis Plot for 25 Participants with the Best Model Fit, Each Graph with Data and Individual Fitted Quadratic Regression Equation

Figure 4.15: Trellis Plot for 25 Participants with the Worst Model Fit, Each Graph with Data and Individual Fitted Quadratic Regression Equation

Figure 4.16: Side-by-Side Normality Q-Q Plots of the Random Effects

Figure 4.17: Multivariate Normality Test for $b_i$

Figure 4.18: The Sampling Distribution of the Estimate of the READ 180 Treatment Effect from the Bootstrap

Figure 4.19: Normality Q-Q Plot of the 1000 Estimates of Treatment Effect based on Independent Bootstrapped Samples
Figure 4.20: Power Curves Produced by the Bootstrap and the SAS PROC MIXED

Figure 4.21: Probability Distributions of the Estimate of the READ 180 Treatment Effect via the Bootstrap and the Monte Carlo with Comparable Design Factors

Figure 4.22: Monte Carlo Simulated Null Treatment Effect Distribution vs. Alternative Treatment Effect Distribution for Sample Size of 300 and 5 Time Points

Figure 4.23: Power Curves Produced by the Bootstrap and the Monte Carlo with Comparable Design Factors

Figure 4.24: Simulated Distributions of the Treatment Effect Estimate at the Null, .025 Quantile, .25 Quantile, Mean, .75 Quantile, and .975 Quantile based on the Bootstrap

Figure 4.25: Power Curves for the Treatment Effect at the .025 Quantile, .25 Quantile, Mean, .75 Quantile, and .975 Quantile at Different Type I Error Rates

Figure 4.26: Power Pattern at the .025 Quantile, .25 Quantile, Mean, .75 Quantile, and .975 Quantile of the Treatment Effect for α at .01 and .05

Figure 4.27: Power Pattern at Different Effect Sizes for α at .01 and .05

Figure 4.28: Simulated Distributions of the Treatment Effect Estimate at the Null and the Bootstrap Estimated Mean with Unbalanced Data at 5 Points in Time for Sample Size of 150, 300, and 600

Figure 4.29: Power Curves for Sample Size of 150, 300, and 600 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated Mean

Figure 4.30: Power Pattern at Sample Size of 150, 300, and 600 for α at .01 and .05

Figure 4.31: Power Curves for Sample Size of 300 and 600 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated .025 Quantile
Figure 4.32: Power Curves for Sample Size of 300 and 600 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated .25 Quantile .................................................................147

Figure 4.33: Distributions of the Treatment Effect Estimate at the Null and the Bootstrap Estimated Mean for Time Points of 5, 6, and 7 with Unbalanced Data .................................................................149

Figure 4.34: Power Curves for Time Points of 5, 6, and 7 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated Mean with Unbalanced Data ..........................................149

Figure 4.35: Power Pattern at Time Points of 5, 6, and 7 for α at .01 and .05 with Unbalanced Data ..............................................................................................................................................150

Figure 4.36: Distributions of the Treatment Effect Estimate at the Null and the Bootstrap Estimated Mean with Balanced and Unbalanced Data .................................................................151

Figure 4.37: Power Curves for Balanced and Unbalanced Data at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated Mean ..............................................................................................................................................152

Figure 4.38: Distributions of the Treatment Effect Estimate at the Null and the Bootstrap Estimated Mean for Time Points of 5, 6, and 7 with Balanced Data and Sample Size of 300 .................................................................................................................154

Figure 4.39: Power Curves for Time Points of 5, 6, and 7 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated Mean with Balanced Data and Sample Size of 300 .................................................................................................................154

Figure 4.40: Power Pattern at Time Points of 5, 6, and 7 for α at .01 and .05 with Balanced Data and Sample Size of 300 .................................................................................................................155

Figure 4.41: Distributions of the Treatment Effect Estimate at the Null and the Bootstrap Estimated Mean for Time Points of 5, 6, and 7 with Balanced Data and Sample Size of 150 .................................................................................................................155

Figure 4.42: Power Curves for Time Points of 5, 6, and 7 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated Mean with Balanced Data and Sample Size of 150 .................................................................................................................156

Figure 4.43: Power Pattern at Time Points of 5, 6, and 7 for α at .01 and .05 with Balanced Data and Sample Size of 150 .................................................................................................................156
Chapter one establishes the research focus of this study. The chapter starts with background information and brings in the major purpose of the study, followed by identifying the research questions and stating the significance of the problems explored. This chapter also includes basic assumptions of the study and the organization of the dissertation.

1.1 Background of the Study

Reading skills are essential for student achievement in all subject areas (Alliance for Excellent Education, 2002; Fleishman, 2004; Just Read, Florida, 2001). Great emphasis has been placed upon improving early literacy, since research supports the critical role of reading skills during early grades in student long-term literacy acquisition and academic success (Bayder, Brooks-Gunn, & Furstenberg, 1993; Butler, Marsh, Sheppard, & Sheppard, 1985; Cunningham & Stanovich, 1997; Hanson & Farrell, 1995; Juel, 1988; No Child Left Behind Act, 2001; Torgesen, 1998). Various early literacy interventions are found to be effective, and today higher percentages of elementary school students are capable of reading proficiently

Such improvement however does not carry over to reading achievement in secondary schools. Recent data from the National Assessment of Educational Progress indicate that around 70 percent of the nation’s eighth graders read below the proficient level and 27 percent read even below the basic level (Lee, Grigg, & Donahue, 2007). In addition, six million of the U.S. adolescents are reading well below grade level (Alliance for Excellent Education, 2002, 2003a). According to the ACT’s national readiness indicator, approximately half of the ACT-tested high school graduates are not adequately prepared for college-level reading (ACT, 2006). In the workplace, poor reading is considered as a major reason for a shortage of competent job candidates faced by most of businesses (Center for Workforce Preparation, 2002).

The prevalent low literacy of adolescents has several far-reaching ramifications for society. One disturbing fact is that poor readers are at greater risk of dropping out of school since they do not possess the necessary literacy skills to catch up with the curriculum (Kamil, 2003; Papalewis, 2005; Snow & Biancarosa, 2003). Over 3,000 students drop out of high school every day, and students reading below grade level are twice as likely to drop out of school as those who are able to read on or above grade level (Alliance for Excellent Education, 2003b; Fleishman, 2004). Furthermore, it is recognized that youth who drop out of school and the high school graduates with low literacy skills have a low chance for employment. Actually about 60 percent of adults who dropped out of high school are not employed (National
Center for Education Statistics, 2001). Those who do find work are most likely to end up with low-paying jobs. Statistics show that dropouts are three times as likely to live in poverty and to receive public assistance as high school graduates, and they earn much less than their peers who continue to earn a college degree (Alliance for Excellent Education, 2002; U.S. Census Bureau, 2002).

What’s worse, adolescents who drop out of school are three and a half times more likely than high school graduates to commit offenses (Coalition for Juvenile Justice, 2001). One key risk factor for juvenile delinquency is poor school performance while a protective factor is above-average status (Hawkins, Smith, & Catalano, 2003; Honig, 2005). Low academic achievement is one of the most powerful predictors of whether youth will engage in maladaptive behaviors like smoking, drinking, using weapons, and attempting suicide (Coalition for Juvenile Justice, 2001; Deshler, Ellis, & Lenz, 1996). It is known that a very high percentage of incarcerated youth have been unsuccessful in school and that the great majority have a significant reading deficit. According to the Coalition for Juvenile Justice (2001), over a third of all incarcerated youth with the median age of 15.5 years old read below the fourth grade level. Typically, juvenile delinquents lag two or more years behind their peers with regard to basic academic skills. In the field of correctional education, there are various programs aiming at development in literacy and other life skills. Studies suggest that effective education programs have a positive influence on the recidivism rate of incarcerated youth. Even though the national recidivism rate for juvenile delinquents is as high as somewhere between 60
to 84 percent, the reoffense rate for participants in successful correctional education programs can be decreased by 20 percent or more (Coalition for Juvenile Justice, 2001; Steurer, Smith & Tracy, 2001). Based on the *National Adult Literacy Survey* in 1992, quality reading instruction programs reduced the recidivism rate of involved incarcerated youth by over 20 percent (National Center for Education Statistics, 1999).

The above factors point to the urgent need for the U.S. to come up with effective strategies to deal with this national reading crisis (Alvermann, 2001; Biancarosa & Snow, 2004). Identifying intensive reading intervention programs that enhance the reading achievement of struggling adolescent readers is imperative. According to the U.S. Department of Education’s Striving Readers Program, struggling readers generally refer to low-performers who read at least two years below grade level (Apthorp & Clark, 2007).

Among various interventions targeting students who read significantly below grade level, READ 180 is “a premier reading intervention program that employs both computer instruction and print materials that are specifically designed for the older, struggling reader” (Fleishman, 2004, p. 6). Published by Scholastic Inc. in 1999, READ 180 is the outcome of collaborative research in both clinical and classroom settings for more than ten years, and is currently implemented in over 6,000 classrooms nationwide (Scholastic, 2005a). As a 90-minute daily intensive reading program, it combines research-based reading practices including computer-assisted learning, whole- and small-group instruction, and independent reading of high interest
books (Scholastic, 2004, 2006a; Taylor, 2006). The READ 180 Instructional Model consists of a 20-minute whole-group instruction, three rotating small-group sessions of 20 minutes each, and a 10-minute whole-group wrap-up (Scholastic, 2006a). READ 180 has several unique features compared to other reading interventions. With the effective use of technology, the program provides differentiated and flexible instruction by tracking individual student progress over time and adjusting the instruction accordingly (Scholastic, 2005a, 2006a). The Scholastic Reading Inventory (SRI), a computer-adaptive assessment that measures student reading levels, provides an evaluation of overall reading achievement and facilitates assignment of program materials to individual students (Scholastic, 1999, 2001). Specifically, students start the program by taking the SRI, and are then matched to appropriate text and placed at the correct level of software activities. Another major ongoing diagnostic assessment tool is the READ 180 Instructional Software, which helps teachers to address individual student needs by identifying strengths and weaknesses in reading (Scholastic, 2005a). Both students and teachers are provided with continuous assessment and immediate feedback during the intervention. It is also noted that the collection of literature offered by READ 180 focuses on topics of interest to adolescent students, and enhances their motivation for reading (Fleishman, 2004). Again, these books are appropriately matched to student reading levels. In addition, READ 180 provides strong ongoing teacher professional development and support (Neal & Kelly, 2002; Scholastic, 2005a).
Since the launch of READ 180, Scholastic Inc. and many schools that have deployed the program have been jointly conducting impact analyses. Numerous studies consistently show that students receiving READ 180 display remarkable growth on different measures of reading comprehension (Goin, Hasselbring, & McAfee, 2004; Interactive, 2002; Papalewis, 2004; Scholastic, 2004, 2006b; White, Haslem, & Hewes, 2005). When there are comparison groups available, students in READ 180 perform significantly better than those in comparison conditions (Fleishman, 2004; Scholastic, 2005a). Moreover, research demonstrates that READ 180 is effective for various adolescent struggling readers, including students with learning disabilities, those receiving special education services, and English-Language Learners (Hewes, Palmer, Haslam, & Mielke, 2006; Palmer, 2003; Scholastic, 2005b, 2006b). In addition to the improvement in reading achievement, more positive attitudes and behaviors, as well as overall higher school achievement are detected for students in READ 180 (Papalewis, 2002).

As appealing as these results appear to be, most of the existing effectiveness studies on READ 180 are limited by methodological constraints. Too often, the research design available is inadequate to provide an accurate assessment of the impact of READ 180 (Pearson & White, 2004). Among the studies, the great majority are correlational or quasi-experimental (Scholastic, 2006b). According to the What Works Clearinghouse (WWC) Evidence Standards (2006), causal conclusions of intervention effectiveness studies can be drawn with strong evidence only if they are conducted with high quality randomized controlled trials or regression
discontinuity designs, and with weaker evidence when they are carried out by quasi-experiment with equated groups. Quite a few studies on READ 180 merely provide insufficient evidence based on the WWC standards. Another methodological inadequacy lies in statistical approaches adopted in many longitudinal studies. Longitudinal data require special statistical methods since the set of observations on the same individual tends to be inter-correlated, and this correlation must be taken into account to draw valid scientific inference (Laird & Ware, 1982). However, many longitudinal impact studies on READ 180 are mainly descriptive or fail to use appropriate statistical techniques. To varying degrees, this also limits the valid inference for its intervention effect. Therefore, it is worth the efforts to further seek convincing evidence for the effectiveness of READ 180 by conducting well-designed studies with proper data analysis.

1.2 Purpose of the Research

Given that large numbers of incarcerated youth are illiterate or only marginally literate, an effective reading program is urgently needed to help them become proficient readers. Although READ 180 seems to be a successful intervention for diverse student populations, there is no literature found for the population of incarcerated youth. Still, READ 180 serves as a promising option to improve the reading performance of low-achieving incarcerated youth. As a result, the first purpose of this study is to investigate if the READ 180 program has a meaningful impact on reading achievement of low-performing incarcerated youth.
To draw causal conclusions, a longitudinal experimental design will be used to evaluate the intervention effect. The reading proficiency and growth of the students in READ 180 and comparison group will be measured using the SRI instrument. In addition to checking whether the mean performances of experimental and comparison populations differ after treatment, the present research concentrates on understanding the dynamics of the learning process and describing how the change takes place. In particular, the study intends to explore the pattern of change across time and also quantify the rate of change over the course of investigation. Meanwhile, this study controls for the influences of salient subject covariates such as age, gender, race/ethnicity, disability status, grade level, school mobility, and other assessments taken at the baseline including Reading and Math of the California Achievement Test (CAT). Mixed-effects modeling will be adopted for analysis due to its important advantages over many other approaches.

The other purpose of this study focuses on characterizing the sampling distribution of the estimate for the treatment effect of READ 180 on low-performing incarcerated youth, and methodologically guiding similar future longitudinal research with power analysis. This in fact concerns how well the intervention effect is modeled. Several simulation studies will be carried out to achieve the purpose. Specifically, the bootstrap technique, a resampling method, will be employed to identify the probability distribution of the treatment effect estimate based on the available sample. The Monte Carlo method is also used to evaluate statistical properties of the estimate for the target population. Furthermore, a series of Monte
Carlo simulations are conducted to examine statistical power for longitudinal research under various circumstances. Defined as the probability of detecting a real treatment effect, power is influenced by many factors such as Type I error rate, effect size, sample size, and error variance (Cohen, 1988; Keppel & Wickens, 2004; Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). In the present study, power is calculated via simulations as well as an analytical approximation in SAS. In addition, power curves will be constructed to show how power differs when changing sample size, number of time points, Type I error rate, magnitude of treatment effect, for both unbalanced and balanced data.

1.3 Research Questions

Based on the two major research purposes, the present study is conducted to address the following questions:

1. Does the READ 180 program have a significant impact on the reading level of low-achieving incarcerated youth over time, when comparing the experimental group with the comparison group being instructed with the traditional educational reading program on reading achievement and growth?

2. Can the variability in the initial reading status of low-performing incarcerated youth be attributed to subject characteristics including age, gender, race/ethnicity, disability status, grade level, special education status, school mobility, and other baseline assessments
including Reading and Math of the California Achievement Test (CAT)?

3. Can the variability in the reading growth rate of low-performing incarcerated youth be accounted for by subject characteristics including age, gender, race/ethnicity, disability status, grade level, special education status, school mobility, and other baseline assessments including Reading and Math of the CAT?

4. What are the statistical properties of the probability distribution of the READ 180 intervention effect estimate, in terms of mean, standard error, and confidence intervals?

5. What is power of the statistical test of the READ 180 intervention effect, based on both the available sample from a pilot study perspective and the simulations?

6. Based on Monte Carlo studies, how does this power change in the context of longitudinal research when varying sample size, number of time points, Type I error rate, magnitude of treatment effect, and data pattern?

1.4 Significance of the Study

This methodological study will contribute to the literature on the READ 180 program by investigating its intervention effect for low-performing incarcerated youth, as there is almost no information available for its impact on this population.
Also given that the study employs a purely experimental design, it meets the WWC standards for strong evidence. This makes the investigation especially unique and informative.

An effective reading program that enhances reading proficiency of incarcerated youth is expected to help these adolescents achieve sufficient literacy skills that are required in school and in life. Identifying such programs is crucial in that juvenile delinquency and recidivism can be reduced with improvement in reading and overall academic achievement, which contributes greatly to public safety and well-being. It is also hoped that with appreciable gains in reading ability these youth can become productive members after their re-entry into society.

Furthermore, relevant findings from simulation studies may shed light on similar longitudinal studies in the future. In particular, the present study is expected to provide useful methodological guidance as it employs an empirically verified model and simulates realistic unbalanced data for power analysis.

1.5 Basic Assumptions

The basic assumptions underlying this study include the following:

1. The READ 180 program is implemented to these low-performing incarcerated youth with high fidelity. Specifically, the instructional model of the daily program is followed according to the specific requirements.
2. The major test instrument used in the study, namely the SRI, has high validity and reliability.

3. Subjects do their best to perform the test. In other words, their test scores accurately reflect their reading proficiency.

4. Information collected for relevant subject covariates are accurate.

5. The model selected for simulation studies fits closely in the target population.

1.6 Organization of the Dissertation

This dissertation consists of five chapters. Chapter one briefly introduces the research questions, the purpose, and the assumptions of this study. Chapter two focuses on comparing different statistical approaches for longitudinal data analysis, as well as reviewing important simulation studies and power analysis. Chapter three describes the target population and study sample, the experimental design, the instrumentation, data collection procedures, data analysis plan, and simulation designs for the study. Chapter four presents the findings for each of the research questions. Chapter five provides a summary of major research findings, discussion of findings with respect to relevant literature, limitations of the study, and recommendations for future research and educational policies.
CHAPTER 2

REVIEW OF THE LITERATURE

Chapter two mainly devotes to methodological issues in analyzing longitudinal or repeated measures data. It first briefly summarizes pertinent literature for the Scholastic READ 180 Instructional Model and salient factors regarding adolescent reading achievement, and then focuses on comparing various statistical approaches to modeling longitudinal data. This chapter also reviews power analysis in mixed-effects modeling and relevant simulation studies.

2.1 The Scholastic READ 180 Instructional Model

According to the Scholastic (2005a), the READ 180 program incorporates the following six critical elements: scientific research base, proven results, comprehensive instruction, purposeful assessment, data-driven instruction, and professional development.

The READ 180 Instructional Model provides a highly structured way to deliver instruction and organize classroom activities (Scholastic, 2004, 2005a). By combining research-based reading practices with the effective use of technology, READ 180 offers adolescent readers an opportunity to experience reading success via
a combination of instructional, modeled, as well as independent reading components (Fleishman, 2004; Scholastic, 2004, Taylor, 2006).

In general, the READ 180 model has the following features (Scholastic, 2004, p. 3-4):

- 90-minute daily class period
- Reduced class size of 15 students per class
- READ 180 Software that provides students with daily, intensive, individualized practice
- Daily modeled or independent reading practice
- Daily individual or small-group instruction
- Whole-group instruction in word analysis, vocabulary development, reading comprehension, and writing
- Distinct classroom areas are designed for each type of instructional activity including, a computer area with five computers for the READ 180 instructional Software, a comfortable reading area with cassette players and headphones for listening to the READ 180 Audiobooks, and a worktable for teacher-directed small-group Instruction.”

The typical READ 180 class consists of four 20-minute blocks and one 10-minute closing wrap-up. The first block introduces students to the lesson and provides opportunity for shared reading, read aloud, or mini skill lessons via whole-group instruction. In the next block, the students are divided into three groups: One group
engages with the teacher in a small group discussion about the introduced lesson. Another group engages in individual computer instruction adapted to their reading levels. The remaining group engages in independent reading of appropriate books. After 20 minutes the groups switch and go to another station, until the 60-minute rotation is complete for three small-group activities. At the end of the READ 180 class, the teacher holds a 10-minute whole-group wrap-up for summary and discussions. (Scholastic, 2004, 2005a)

2.2 Salient Factors regarding Adolescent Reading Achievement

Researchers acknowledge that limited information is available to understand adolescent reading (Hasselbring, & Goin, 2004; National Reading Panel, 2000; Pressley, 2000). Still, studies consistently indicate that race/ethnicity and poverty/economic status are salient factors influencing students’ reading abilities (Daggett & Hasselbring, 2007). Disproportionately large numbers of non-White students are at risk, and the reading achievement gap between minority students and White has been discussed extensively (Jencks & Phillips, 1998). Although this gap may narrow somewhat (National Center for Education Statistics, 2002), it is acknowledged that minorities are still at a disadvantage (McCoach, O’Connell, Reis, & Levitt, 2006). While around 41 percent of White eighth graders read at or above the proficient level, merely 13 percent of African American and Hispanic students in the eighth grade belong to the proficient category (Grigg, Daane, Jin, & Campbell, 2002). Generally the below-basic achievement group is estimated to be twice as large.
among African American and Hispanic students as in other groups (Daggett & Hasselbring, 2007).

Moreover, poverty/economic status is widely acknowledged to be a very important predictor of academic achievement in the U.S. today (Smith, Brooks-gunn, & Klebanov, 1997). Studies consistently support the relationship between family SES and reading achievement (Molfese, Modglin, & Molfese, 2003). There seems to be a strong connection between family poverty and the increased possibility of low literacy (Daggett & Hasselbring, 2007; Papalewis, 2005). In fact, approximately 68 percent of the poorest students in the fourth grade failed to accomplish basic levels of literacy (Donahue, Voelkl, Campbell, & Mazzeo, 1999). Furthermore, SES seems to be connected to changes in students’ achievement trajectories as they advance through school (Jimerson, Egeland, & Teo, 1999).

Gender and disability status may serve as another two salient factors. Generally female students outperform their male counterparts somewhat in reading, with a higher percentage of girls achieving reading proficiency (National Center for Education Statistics, 2002). With regard to disability, it is estimated that approximately one third of the students with various disabilities in the U.S. lack reading proficiency and therefore in need of special education services (U.S. Department of Education, 2002).
2.3 Statistical Approaches for Longitudinal Data Analysis

The statistical analysis of longitudinal/repeated measures data over time can be a remarkably challenging task, with the potential to throw light on many important theoretical questions of interest. Longitudinal/repeated measures studies generally aim at characterizing patterns of the subjects’ response over time and defining relationships between these patterns and covariates (Ware, 1985). Sometimes these studies may simply focus on change in a dependent variable for subjects over time. There are a lot of longitudinal data in various disciplines such as social sciences, medical sciences, economics, agriculture and industry. To date, a wide variety of longitudinal statistical models have been proposed to address the challenge, with the literature being expanded rapidly in the past two decades (see Bryk & Raudenbush, 1987, 1992; Goldstein, 1987, 1995, 2003; Hox, 1994, 1995, 2000; Lindsey, 1993; Littell, Milliken, Stroup, & Wolfinger, 1996). Comparisons between different models are made in the following sections.

2.3.1 Repeated Measures ANOVA

Among all the available models for repeated measures data, the Analysis of Variance (ANOVA) is one of the most commonly used methods. Under this approach, the outcome variable for a subject is decomposed into the effects associated with “between-subject” factors and “within-subject” factors (Dean & Voss, 1999; Jennrich & Schluchter, 1986; Keppel & Wickens, 2004; Montgomery, 2005).
Consider a typical study with $N$ subjects that belong to $G$ treatment groups.

The model for the repeated measures ANOVA can be written as follows:

$$y_{igj} = \mu + \tau_g + \gamma_j + (\tau\gamma)_{ig} + b_{ig} + e_{igj} \quad 1 \leq g \leq G; 1 \leq j \leq n; 1 \leq i \leq N_g$$  \hspace{1cm} (2.1)

where $\mu$ is the overall mean, $\tau_g$ is the effect associated with “between-subject” factors for group $g$, $\gamma_j$ is the effect associated with “within-subject” factors at $j^{th}$ measurement occasion, $(\tau\gamma)_{ig}$ is the interaction effect for group $g$ at occasion $j$, $b_{ig}$ is the random effect for subject $j$ in group $g$, $e_{igj}$ is the random effect for subject $i$ in group $g$ at $j^{th}$ measurement occasion, and $N_g$ is the number of subjects in group $g$.

Random effects $b_{ig}$ are random over individuals and assumed to follow a normal distribution with zero mean, i.e.,

$$b_{ig} \sim N(0, \varphi).$$  \hspace{1cm} (2.2)

Random effects $e_{igj}$ are random over the repeated measurements within an individual and assumed to follow a normal distribution with zero mean, i.e.,

$$e_{igj} \sim N(0, \sigma^2).$$  \hspace{1cm} (2.3)

The distribution of $n$ repeated measurements for subject $j$ in group $g$ is as follows:

$$y_{ig} \sim N(\mu_{ig}, \Sigma)$$  \hspace{1cm} (2.4)

where

$$\mu_{ig} = \left[ \begin{array}{c} \mu + \tau_g + \gamma_1 + (\tau\gamma)_{g,1} \\ \mu + \tau_g + \gamma_2 + (\tau\gamma)_{g,2} \\ \vdots \\ \mu + \tau_g + \gamma_n + (\tau\gamma)_{g,n} \end{array} \right]_{nx1}$$  \hspace{1cm} (2.5)
and

\[
\Sigma = \text{cov}(y_{ig}) = \begin{bmatrix}
\varphi + \sigma^2 & \varphi & \cdots & \varphi \\
\varphi & \varphi + \sigma^2 & \cdots & \varphi \\
\vdots & \vdots & \ddots & \vdots \\
\varphi & \varphi & \cdots & \varphi + \sigma^2
\end{bmatrix}_{\times n}.
\] (2.6)

Therefore, the variance-covariance matrix of \( n \) measurements for each subject is assumed to be the same for all subjects and has a fixed “compound symmetry” pattern as shown in the above equation (Keppel & Wickens, 2004).

Although the repeated measures ANOVA is a classical approach, it suffers some major shortcomings. First, the model is not flexible and encounters difficulties if studies do not have regularly and consistently timed data (Wallace & Green, 2002). Basically it is suitable for the balanced repeated measures data, i.e., the measurements on each subject are taken at exactly the same occasions without any missing data (Maas, & Snijders, 2003). Thus this approach is somewhat limited since unbalanced data in longitudinal studies are quite normal. Furthermore, different times of measurements cannot be directly incorporated in the analysis (Cudeck, 2007).

Second, the model does not explicitly incorporate the actual time when the repeated measures are taken for the subjects. Instead, the time variable is treated as a discrete variable rather than a continuous one. Two separate fixed effect terms, \( \gamma_j \) and \( (\tau_j)_{ig} \), are used to associate with each occasion. It this way, it does not take into any account of the time variable directly and lacks important information about the time of the measurements (Cudeck, 2007). Third, a uniform structure of the mean vector is imposed by the model, and this lacks flexibility for some repeated measures
experiments (Cudeck, 2007). Fourth, the variance-covariance structure of the measurement vector for each subject always has a compound symmetry structure (Keppel & Wickens, 2004). This restrictive structure does not seem to be suitable for most of longitudinal or repeated measures studies and is considered to be another drawback of the repeated measures ANOVA model (Cudeck, 2007). Although the multivariate Repeated Measures ANOVA (MANOVA), in which a set of measures for a subject are treated as a collection of multivariate scores instead of a series of univariate scores, can be adopted to give flexibility to the variance-covariance structure of the measurement vector of a subject, this model runs into another extreme where the variance-covariance structure is always unrestricted and no available simpler structure could be imposed by the model (Bock, 1985; Johnson & Wichern, 1988; Morrison, 1990). This causes additional coefficients for estimation and negatively affects the precision of the entire model especially when the sample size is small (Cudeck, 2007).

2.3.2 Extended Linear Model for Repeated Measures

The extended linear model is developed based on the linear model (Crowder & Hand, 1993; Cudeck, 2007; Hand & Taylor, 1987). It can be written as

\[ y_i = X_i \beta + e_i \]  \hspace{1cm} (2.7)

where the matrix of known variates \( X_i \) changes for each subject according to their group, the vector of fixed \( \beta \) includes regression coefficients and fixed effects for
“between-subject” factors, e.g., treatment effect, and $e_i$ is the vector of residuals following a multivariate normal distribution with zero means

$$e_i \sim N(0, \Sigma_j). \quad (2.8)$$

The main difference between an extended linear model and a traditional linear model is that the $X_i$ matrix is set according to subjects’ group. For instance, consider a longitudinal study comparing two groups in which a linear regression model is assumed:

$$y_{ij} = \begin{cases} 
\beta_0 + \beta_1 t_j + e_{ij} & \text{subject } i \text{ in the control group} \\
\beta_0 + (\beta_1 + \delta) t_j + e_{ij} & \text{subject } i \text{ in the treatment group}
\end{cases} \quad (2.9)$$

where $\beta_0$ is the initial status of a subject at time $t = 0$, $\beta_1$ is the rate of improvement for a subject in the control group, and $\delta$ is the treatment effect for the rate of improvement. With this design, the $X_i$ matrix of the extended linear model can be constructed as follows:

$$X_i = \begin{bmatrix} 
1 & t_1 & g_i t_1 \\
1 & t_2 & g_i t_2 \\
\vdots & \vdots & \vdots \\
1 & t_n & g_i t_n 
\end{bmatrix} \quad (2.10)$$

where $g_i$ equals 0 for a subject $i$ in the control group and equals 1 for one in the treatment group. The vector $\beta$ can be constructed as

$$\beta = [\beta_0, \beta_1, \delta]^\prime. \quad (2.11)$$

In the extended linear model, the distribution of the observation variables for subject $i$ is

$$y_i \sim N(X_i \beta, \Sigma_j). \quad (2.12)$$
This extended linear model is built based on the linear model. It allows comparing results between different groups and testing the treatment effect. It also has the flexibility for the variance-covariance structure of the dependent variables, $\Sigma_e$, for a subject. Various variance-covariance structures can be applied to the extended linear model. However, given that the model assumes that all subjects within a group follow a regression model with the same set of regression parameters, no random effects for regression coefficients over individual subjects are allowed in the model.

2.3.3 Random Coefficient Model for Repeated Measures

The extended linear model can be naturally further extended to allow random effects for regression coefficients over individual subjects. This type of model is usually called the random coefficient model and can be written as follows (Hand & Crowder, 1996; Longford, 1993; Zegar, Liang, & Albert, 1988):

$$
\mathbf{y}_i = \mathbf{X}_i (\mathbf{\beta} + \mathbf{b}_i) + \mathbf{e}_i \\
= \mathbf{X}_i \mathbf{\beta} + \mathbf{X}_i \mathbf{b}_i + \mathbf{e}_i
$$

(2.13)

The underlying philosophy of a random coefficient model is that each variable always has a deterministic part and a stochastic part. In a random coefficient model, each regression coefficient or treatment effect for different groups will be associated with a random effect. With regard to the example described in the section of extended linear model, its random coefficient model expanded from the extended linear model in non-matrix format can be written as
\[ y_{ij} = \begin{cases} 
(\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_j + e_{ij} & \text{subject } i \text{ in the control group} \\
(\beta_0 + b_{i0}) + (\beta_1 + b_{i1} + \delta + b_{i2})t_j + e_{ij} & \text{subject } i \text{ in the treatment group} 
\end{cases} \] (2.14)

and matrix \( X_i \) and vector \( \beta \) are the same as those for the extended linear model. The vector \( b_i \) can be constructed as follows:

\[ b_i = [b_{i0}, b_{i1}, b_{i2}]'. \] (2.15)

The vector \( b_i \) is assumed to follow a normal distribution with zero means, i.e.,

\[ b_i \sim N(0, \Phi). \] (2.16)

Therefore, the dependent variables of subject \( i \) follows the following distribution

\[ y_i \sim N(X_i\beta, X_i\Phi X_i' + \Sigma). \] (2.17)

Compared to the extended linear models, random coefficient models have several characteristics regarding the variance-covariance structure (Cudeck, 2007). First, the variance-covariance matrix of \( y_i \) differs from individual to individual as the response model does. Second, the variance-covariance matrix of \( y_i \) also depends on the matrix \( X_i \) that includes independent variables of time of measurements. In this way, information about the time of measurements for subjects is also incorporated into the variance-covariance structure. Third, it adds the flexibility to allow the variance-covariance matrix of \( y_i \) to differ from the variance-covariance matrix of the residuals \( e_i \). This allows one to construct some interesting relationships in the variance-covariance patterns. Fourth, the parameterization of the variance-covariance matrix of \( y_i \) can be kept parsimonious even if the sample size is large. (Cudeck, 2007)
The random coefficient model allows random effects, $b_i$, over individual subjects for regression coefficients and treatment effects (Laird & Ware, 1982). It gained growing interest in many scientific areas. The framework of the model is subject-specific and it associates a random effect with each fixed effect for regression coefficient or treatment effect (Cudeck, 2007). Therefore, it has a lack of flexibility for constructing random effects models and fixed effects models.

2.3.4 Mixed-Effects Model for Repeated Measures

By uncoupling the fixed and random effects in a random coefficient model, the flexibility of constructing the random effects and fixed effects models can be added. This is motivated by the fact that some of stochastic parameters in the random coefficient models can be zero (Cudeck, 2007). Therefore, a mixed-effects model for the study in the previous random coefficient model can be written as follows:

$$y_i = X_i \beta + Z_i b_i + e_i.$$  \hspace{1cm} (2.18)

This model is more elaborate in that it includes the standard random effects model as a special case. In a special case where all fixed effects are associated with random effects, the model becomes a random coefficient model and $X_i = Z_i$. However, the mixed-effects model has the flexibility that if one or more of the random effects is unnecessary those coefficients can be eliminated from the model. (Cudeck, 2007; Wallace & Green, 2002) For the model described previously, when a random effect for the fixed treatment effect $\delta$ is not necessary, a mixed-effects model can be constructed as follows:
\[
y_{ij} = \begin{cases} 
(\beta_0 + b_{10}) + (\beta_1 + b_{11})t_j + e_{ij} & \text{subject } i \text{ in the control group} \\
(\beta_0 + b_{10}) + (\beta_1 + \delta + b_{11})t_j + e_{ij} & \text{subject } i \text{ in the treatment group}
\end{cases}
\] (2.19)

The matrix \( Z_i \) can be constructed as follows:
\[
Z_i = \begin{bmatrix}
1 & t_1 \\
1 & t_2 \\
\vdots & \vdots \\
1 & t_n
\end{bmatrix}
\] (2.20)

and the vector \( b_i \) becomes
\[
b_i = [b_{10}, b_{11}]'.
\] (2.21)

The dependent variables of subject \( i \) follows the following distribution
\[
y_i \sim N(X_i\beta, Z_i\Phi Z_i' + \Sigma_i).
\] (2.22)

Given this flexibility of the mixed-effects model, it can be treated as a two-level covariates model and constructed by use of a two-stage analysis. In this way, the fixed effects due to various between-subject factors can be incorporated in the model to understand and explain between-subject differences in the regression coefficients for corresponding covariates. At the same time, the random effects model can be kept simple. This highly facilitates the model fitting process and model interpretation. (Verbeke & Molenberghs, 1997, 2000)

The mixed-effects model has several important features for describing longitudinal data (Cudeck, 2007). First, both inter-individual and intra-individual variability is specified explicitly in the model, which allows researchers to model both sources in an appropriate way in various applications. Second, mixed-effects models relate the fixed and random parts in a single package and also incorporate a
sensible structure of the population mean trajectory. Third, the mixed-effects models allow researchers to include various types of covariates in the second stage analysis, which provide more information to moderate or enhance the understanding of the change process in longitudinal studies and also extend the flexibility of the model. In addition, one should bear in mind that both the extended linear model and the random coefficient model are special cases of the mixed effects model. (Cudeck, 2007; Verbeke & Molenberghs, 1997, 2000)

Therefore, a mixed-effects model is a preferable approach for complex longitudinal data analysis. The mixed-effects models constructed by a two-stage analysis for the present study are further described in details in following sections.

2.3.4.1 Covariance Structures for Residuals

With regard to different models for analyzing cross-sectional data, the structure of the regression residuals typically remains the same: the residuals have zero mean and constant variance, and are independent between individuals (Kutner, Nachtsheim & Neter, 2004). When applying regression models to longitudinal/repeated measures data, however, this assumption may not necessarily holds due to the following reasons: 1). Residuals for a given individual tend to correlated. 2). Residuals normally become more or less variable over time instead of staying constant (Cudeck, 2007; Ware, 1985).

Therefore, different covariance structures for residuals can be compared when fitting a mixed-effects model to longitudinal data. The commonly used structures for
residuals in mixed-effects modeling are described in the following (Crowder & Hand, 1993; Fitzmaurice, Laird, & Ware, 2004; Verbeke & Molenberghs, 1997, 2000).

When there are 3 repeated measures, an unstructured covariance structure allows an arbitrary pattern of variances and covariances:

\[
\begin{bmatrix}
\sigma_1^2 \\
\sigma_{12} & \sigma_2^2 \\
\sigma_{13} & \sigma_{23} & \sigma_3^2 \\
\end{bmatrix}
\]

In sharp contrast, a simple structure based on the aforementioned sphericity assumption would be as follows:

\[
\begin{bmatrix}
\sigma_1^2 \\
0 & \sigma_2^2 \\
0 & 0 & \sigma_3^2 \\
\end{bmatrix}
\]

As a generalization of the simple structure, the following banded (1) pattern is also called heterogeneous variances:

\[
\begin{bmatrix}
\sigma_1^2 \\
0 & \sigma_2^2 \\
0 & 0 & \sigma_i^2 \\
\end{bmatrix}
\]

A more general banded (2) pattern also allows correlation between adjacent measures:

\[
\begin{bmatrix}
\sigma_1^2 \\
\sigma_{12} & \sigma_2^2 \\
0 & \sigma_{23} & \sigma_3^2 \\
\end{bmatrix}
\]

The compound symmetry structure is shown as follows:

\[
\begin{bmatrix}
\sigma_1^2 + \sigma_i^2 \\
\sigma_i^2 & \sigma_1^2 + \sigma_i^2 \\
\sigma_i^2 & \sigma_i^2 & \sigma_1^2 + \sigma_i^2 \\
\end{bmatrix}
\]
In the first-order autoregressive correlation pattern, pairs of variables separated by the same number of steps are correlated equally, with the correlation in each diagonal band being less than or equal to bands above it while being greater than or equal to elements in bands below it (Cudeck, 2007). This structure is shown below:

\[
\begin{bmatrix}
\sigma^2 & \rho \sigma^2 & \rho^2 \sigma^2 \\
\rho \sigma^2 & \sigma^2 & \rho^2 \sigma^2 \\
\rho^2 \sigma^2 & \rho^2 \sigma^2 & \sigma^2
\end{bmatrix}
\]

The Toeplitz structure is based on the assumption that the variance is constant across occasions and also that any pair of responses that are equally separated in time have the same correlation (Fitzmaurice et al., 2004). The pattern is shown as follows:

\[
\begin{bmatrix}
\sigma^2 & \sigma_{12} & \sigma_{13} \\
\sigma_{12} & \sigma^2 & \sigma_{12} \\
\sigma_{13} & \sigma_{12} & \sigma^2
\end{bmatrix}
\]

2.3.4.2 Estimation Approaches

2.3.4.2.1 Maximum Likelihood Estimation (ML)

In the mixed-effects model defined in Section 2.3.4,

\[
\begin{align*}
Y_i = X_i \beta + Z_i \beta_i + e_i \\
\beta_i & \sim N(0, \Phi) \\
e_i & \sim N(0, \Sigma) \\
\beta_i, e_i & \text{ independent}
\end{align*}
\]
Consequently, $Y_i$ is normally distributed with a mean of $X_i\beta$ and a variance of $V_i = Z_i\Phi Z_i' + \Sigma_i$, and the mixed-effects model implies the following marginal model (Verbeke & Molenberghs, 2000):

$$Y_i \sim N(X_i\beta, Z_i\Phi Z_i' + \Sigma_i)$$

(2.24)

Let $\alpha$ denote the vector of all variance and covariance parameters in the above model. Therefore, it consists of $q(q+1)/2+m$ different elements in $\Phi$ and $\Sigma_i$. Also let $\theta = \begin{bmatrix} \beta \\ \alpha \end{bmatrix}$ represent the $s$-dimensional vector of all parameters involved in the mixed-effects model for $Y_i$. The classical approach to inference is based on the estimates obtained from maximizing the likelihood function specified as

$$L_{ML}(\theta) = \prod_{i=1}^{n} \left\{ (2\pi)^{-n/2} |V_i(\alpha)|^{-1/2} \times \exp\left( -\frac{1}{2} (Y_i - X_i\beta)'V_i^{-1}(\alpha)(Y_i - X_i\beta) \right) \right\}$$

(2.25)

with respect to the parameter vector $\theta$.

The maximum likelihood estimator (MLE) of $\beta$, obtained from maximizing (2.25), conditional on $\alpha$, i.e., assuming $\alpha$ is known, is given by

$$\hat{\beta}(\alpha) = \left( \sum_{i=1}^{n} X_i'V_i^{-1}(\alpha)X_i \right)^{-1} \sum_{i=1}^{n} X_i'V_i^{-1}(\alpha)y_i.$$  

(2.26)

As a result, the MLE of $\alpha$ is obtained by maximizing

$$L_{ML}(\theta) = \prod_{i=1}^{n} \left\{ (2\pi)^{-n/2} |V_i(\alpha)|^{-1/2} \times \exp\left( -\frac{1}{2} (Y_i - X_i\hat{\beta}(\alpha))'V_i^{-1}(\alpha)(Y_i - X_i\hat{\beta}(\alpha)) \right) \right\}$$

(2.27)

with respect to $\alpha$. Then the MLE of $\beta$ can be calculated via (2.26) by substituting $\alpha$ with $\hat{\alpha}$. (Verbeke & Molenberghs, 2000)
2.3.4.2.2 Restricted Maximum Likelihood Estimation (REML)

The mixed-effects model described previously in (2.23) for all subjects can be combined in the following model:

$$Y = X\beta + Zb + e$$

(2.28)

where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \quad Z = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}, \text{ and } e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

The dimension of $Y$ is $(N \times N)$, where $N = \sum_{i=1}^{n} N_i$. Therefore,

$$Y \sim N(X\beta, V(\alpha))$$

(2.29)

where

$$V(\alpha) = \begin{bmatrix} V_1(\alpha) & 0 & \cdots & 0 \\ 0 & V_2(\alpha) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_n(\alpha) \end{bmatrix}.$$  

(2.30)

The REML estimator for the variance components $\alpha$ can be obtained by maximizing the likelihood function of a set of error contrasts $U = A'Y$ where $A$ is any $(n \times (n-p))$ full-rank matrix with the columns orthogonal to the columns of the $X$ matrix, i.e., $A'X = 0$. The vector $U$ then follows a normal distribution with zero mean vector and the covariance matrix $A'V(\alpha)A$, which does not depend on $\beta$. The likelihood function of the error contrasts can be constructed as
\[ L(\alpha) = (2\pi)^{-(n-p)/2} \left| \sum_{i=1}^{n} X_i'X_i \right|^{1/2} \]
\[ \times \sum_{i=1}^{n} X_i'V^{-1}X_i \left| \prod_{i=1}^{n} |V_i| \right|^{1/2} \]
\[ \times \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} (Y_i - X_i\hat{\beta})'V^{-1}(Y_i - X_i\hat{\beta}) \right\} \]  

(2.31)

where \( \hat{\beta} \) is given by (2.26). Consequently, the REML estimate \( \hat{\alpha} \) does not depend on the error contrasts (i.e., the choice of \( A \)). And the likelihood function in (2.31) is equivalent to

\[ L(\alpha) = C \left| \sum_{i=1}^{n} X_i'V^{-1}(\alpha)X_i \right|^{1/2} L_{ML} \left( \hat{\beta}(\alpha), \alpha \right) \]  

(2.32)

where \( C \) is a constant independent of \( \alpha \), and \( L_{ML} \left( \hat{\beta}(\alpha), \alpha \right) = L_{ML} (\theta) \) is the maximum likelihood function given in (2.27).

Since \( \left| \sum_{i=1}^{n} X_i'V^{-1}(\alpha)X_i \right| \) in (2.31) does not depend on \( \beta \), it follows that the REML estimates for \( \alpha \) and for \( \beta \) can also be obtained by maximizing the so-called REML likelihood function

\[ L_{REML} (\theta) = \left| \sum_{i=1}^{n} X_i'V^{-1}(\alpha)X_i \right|^{1/2} L_{ML} (\theta) \]  

(2.33)

with respect to all parameters simultaneously (\( \alpha \) and \( \beta \)). (Verbeke & Molenberghs, 2000)
2.3.4.2.3 A Comparison of ML and REML

As described in the previous sections, in the ML fitting process, the likelihood of the whole sample data is maximized to give the estimates of the fixed effects and the variance-covariance matrices of random effects. Thus the estimates of the variance and covariance matrices of random effects are conditional on the point estimates of the fixed effects for ML (Raudenbush & Bryk, 2002). In contrast, REML maximizes the likelihood of the level-1 residuals and the level-2 random effects (Singer & Willett, 2003). The number of parameters for REML is equal to the number of the variance and covariance components of the random effects, whereas the number of parameters for ML is the sum of the number of fixed effects and the number of the variance and covariance components of the random effects (McCoach & Black, 2008).

The two methods typically yield similar estimates for the variance and covariance components under the condition that the number of level-two units, i.e., the number of individual subjects in longitudinal studies, is large. However, when that number of subjects is relatively small, ML and REML can produce different results and normally the REML estimates of the variance-covariance components of the random effects are considered to be more realistic (Raudenbush & Bryk, 2002). This is due to the fact that ML does not adjust the estimates of the variance and covariance matrices of random effects for the uncertainty about the fixed effects (McCoach & Black, 2008).
2.3.4.3 Evaluation of Model Fit and Model Selection Criteria

Generally when choosing a final model among candidate models, one should consider the interpretability, the fit indices, and the principle of parsimony (Burnham & Anderson, 2004; Myung, 2000).

There are two most commonly used approaches of model selection criteria. One is the hypothesis test, and the other is the “information criteria” approach (McCoach & Black, 2008). They are both based on the concept of deviance of a model.

As discussed previously, maximum likelihood estimation is used to provide estimates for the fixed effects and the variance-covariance components of random effects in mixed-effects models. In the ML estimation, a likelihood function describes the probability of obtaining the observed data given that the unknown parameters of the model and the parameter estimates are those that give the maximum value of the likelihood function (Singer & Willett, 2003). Therefore, this maximum likelihood is also available in addition to the parameter estimates in the ML estimation. A deviance statistic of the model can therefore be obtained from this likelihood for the model (Snijders & Bosker, 1999).

A deviance statistic of a model is defined as

\[ D = -2\left[\log(\text{likelihood of current model}) - \log(\text{likelihood of saturated model})\right] \] (2.34)

The deviance compares the logarithm of the likelihood of the current model to that of a saturated model (Singer & Willett, 2003). The saturated model fits the observed data perfectly and is the best possible model for the sample of data at hand, and thus
the deviance is a measure of how much worse that the model is as compared to the
saturated model (Singer & Willett, 2003). Therefore, the deviance statistic is relative
to the saturated model and dependent on the sample size and the fit of the model, and
it cannot be interpreted directly. However, the difference in deviance between
multiple competing models can be interpreted when they are hierarchically nested
models, and the same observed data set is fitted to the models by use of the ML
estimation approach (McCoach & Black, 2008).

The hypothesis test is one of the most common model selection methods
(Weaklim, 2004). A chi-square likelihood ratio test can normally be used to compare
the fitness of the two competing models in the mixed-effects modeling.

The difference between the deviances of the simpler model and that of the
more complex model is defined as follows:

\[ \Delta D = D_1 - D_2 \]  

(2.35)

It describes the change in deviance for the two models. The difference of the
deviance statistics of two models can be compared directly given that the two models
are hierarchically nested (McCoach & Black, 2008).

In general, a more complex model with a greater number of parameters has
better fit and lower deviance than that of the simpler one. For sample data with large
size, the difference calculated in (2.35) between the two models follows an
approximate chi-square distribution with degrees of freedom described in the
following equation (de Leeuw, 2004):

\[ DF = p_2 - p_1 \]  

(2.36)
where $p_1$ and $p_2$ are the number of parameters for the simpler model and the more complex one respectively.

The likelihood ratio test is used to determine if the more complex model with a larger number of parameters reduce the deviance by an amount that is significant enough to reject the simpler model. In practice, the difference in the deviance between the two models is statistically significant and the simpler model is rejected if the change in deviance ($\Delta D = D_1 - D_2$) is larger than the critical value of the chi-square distribution with degrees of freedom of $DF = p_2 - p_1$. Otherwise, the more parsimonious model with a smaller number of parameters will be retained. (McCoach & Black, 2008)

Due to the fact that the ML estimation method includes both the fixed effects and variance-covariance components of random effects in the number of reported parameters whereas REML only counts the number variance-covariance components of random effects in the number of reported parameters, it is necessary to use the ML estimation approach to compare two nested models with different sets of fixed effects (Snijders & Bosker, 1999). REML can only be used to compare two nested models with the same set of fixed effects and different sets of random effects in the chi-square hypothesis test for model selection (McCoach & Black, 2008).

The likelihood ratio test suffers several main drawbacks. First, the hypothesis test approach cannot be used to compare two models that are not hierarchically nested (Raftery, 1995). Second, it provides little guidance for selecting an imperfect but good enough and parsimonious model, and it does not help one in determining if the
lack of fit for a parsimonious model is problematic in practice (Gelman & Rubin, 1995). In addition, it does not quantify how much worse a parsimonious model is when compared with a more complex model (McCoach & Black, 2008). Under these circumstances, the model selection index approaches can be very informative in that they allow the comparison of models with different parameters and the quantification of the degree to which a model has an improvement over others (McCoach & Black, 2008).

There are two commonly used fit indices, namely the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These two index approaches allow one to compare models that are not necessarily nested (Wagenmakers & Farrell, 2004; Wasserman, 2000). They also quantify the magnitude in which a model is better or worse than the other ones (Kuha, 2004).

The AIC index is calculated by the following equation (Akaike, 1973, 1974, 1977, 1985):

\[
AIC = D + 2p
\]  

(2.37)

where \(D\) is deviance of the model and \(p\) is the number of parameters estimated in the model. There are two terms in the formula for calculating the AIC index. The first the deviance statistic of the model and it stands for the degree of badness of fit of the model (Bozdogan, 1987, 2000). The second term is equal to two times the number of parameters of the model and serves as a penalty to the more complex models. Given this penalty, a more complex model has to give an improvement of deviance in a magnitude more than 2 per additional parameters than the simpler model in order to
reject the simpler model. The model with the lowest AIC index among all candidate models is considered to be the best one using this approach. The AIC index normally gives similar results to the chi-square hypothesis test (McCoach & Black, 2008).

The BIC index is calculated by the following equation (Schwarz, 1978; Wasserman, 2000):

\[
\text{BIC} = D + \ln(n) \times p
\]

(2.38)

where \( D \) is the deviance of the model, \( p \) is the number of parameters estimated in the model, and \( n \) is the size of the sample data. Compared to the AIC index, the BIC index imposes a penalty to the models with more parameters which is impacted by the sample size. In this way, the penalty on the number of parameters will be very large when the sample size is large. Due to the fact that BIC explicitly takes sample size into account, it normally favors the most parsimonious models among the three model selection methods described so far (McCoach & Black, 2008).

2.3.4.4 Inferences for Fixed Effects

As discussed in the previous section, the vector \( \mathbf{\beta} \) of fixed effects is estimated by (2.26), in which the unknown vector \( \mathbf{\alpha} \) is replaced with its ML or REML estimate. Furthermore, conditional on \( \mathbf{\alpha} \), the estimate of \( \mathbf{\beta} \) follows a multivariate normal distribution with the mean vector \( \mathbf{\beta} \) and the variance matrix (Verbeke & Molenberghs, 2000):

\[
\text{var}(\hat{\mathbf{\beta}}) = \left( \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{V}^{-1}(\mathbf{\alpha}) \mathbf{X}_i \right)^{-1}
\]

(2.39)
Approximate t-tests and F-tests

For each parameter in $\beta$, an approximate $t$-test and associated confidence interval can be obtained from approximating the distribution of $\frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)}$ by an appropriate $t$-distribution. Hence testing the general linear hypotheses

$$H_0: L\beta = 0 \quad \text{vs.} \quad H_a: L\beta \neq 0 \quad (2.40)$$

can now be based on an $F$-approximation to the distribution of

$$F = \frac{(\hat{\beta} - \beta)'L' \left[ L \left( \sum_{i=1}^{n} X_i'V_i^{-1}(\alpha)X_i \right)^{-1} L' \right]^{-1} L(\hat{\beta} - \beta)}{\text{rank}[L]} \quad (2.41)$$

The numerator degrees of freedom equals $\text{rank}[L]$, while the denominator degrees of freedom needs to be estimated from the actual data. The same is true for the degrees of freedom needed in the $t$-approximation discussed above. (Verbeke & Molenberghs, 2000)

Several methods can be used for estimating the appropriate number of degrees of freedom needed for a specific $t$- or $F$-test. The SAS PROC MIXED procedure includes four different estimation methods (Littell et al., 1996, 2006; Verbeke & Molenberghs, 1997, 2000). S-PLUS also provides some relevant estimation approaches (Pinheiro & Bates, 2000).

Effect size

Regarding mixed-effects models, the effect size index $f^2$ (Cohen, 1988) can be used to measure the size of a fixed effect in a model, which is defined as
$$f^2 = \frac{\text{PV}_s}{\text{PV}_e} = F \frac{u}{v}$$  \hspace{1cm} (2.42)$$

where $\text{PV}_s$ is the proportion of $Y$ variance accounted for by the effect (source), $\text{PV}_e$ is the proportion of error or residual variance, $F$ is calculated by (2.41), and $u$ and $v$ are numerator and denominator degrees of freedom for the noncentral $F$ distribution and can be calculated by

$$\begin{cases} 
    u = \text{rank}[L] \\
    v = \sum_{i=1}^{n} N_i^+ - \text{rank}[X \mid Z]
\end{cases} \hspace{1cm} (2.43)$$

The partial squared correlation ratio $R^2$ is implied by any given value of $f^2$

$$R^2 = \frac{f^2}{1 + f^2} \hspace{1cm} (2.44)$$

and represents the portion of the $Y$ variance accounted for by the effect.

*Small effect size: $f^2 = 0.02$. According to Cohen (1988), a small effect size index $f^2$ is set to be 0.02. Translated into a partial squared correlation, it gives

$$\frac{0.02}{1 + 0.02} = 0.02.$$ 
Thus a small effect accounts for 2% of the $Y$ variance as defined here.

*Medium effect size: $f^2 = 0.15$. Similarly, an effect that accounts for

$$\frac{0.15}{1 + 0.15} = 13\%$$

of the $Y$ variance is defined as a medium effect.

*Large effect size: $f^2 = 0.35$. An effect that accounts for

$$\frac{0.35}{1 + 0.35} = 26\%$$

of the $Y$ variance is defined as a large effect.
2.4 Power Analysis for Mixed-Effects Models

In the field of mixed-effects modeling, power analysis can be carried out by analytical approximation or via simulations, with the latter being the more recommended approach (Kreft, 1995; Castelloe & O’Brien, 2000).

2.4.1 An Analytical Approximation

To test the general linear hypotheses in (2.40), the distribution of $F$, which is defined in (2.41), can also be approximated by an $F$-distribution under the alternative hypothesis $H_a$, with rank[$L$] and $\sum_{i=1}^{N_i} N_i^+ - \text{rank}[X \mid Z]$ degrees of freedom and with the noncentrality parameter

$$\delta = (L\beta)' \left[ L \left( \sum_{i=1}^{n} X_i ' V_i^{-1} (\mu X_i) - L \right) \right]^{-1} (L\beta)$$

(2.45)

Therefore, a noncentral $F$-distribution is obtained under $H_a$, and power calculations immediately follow based on this distribution (Littell et al., 1996; Verbeke & Molenberghs, 2000). This analytical approach is currently incorporated in the SAS program (Littell et al., 2006).

2.4.2 Simulation Studies on Power

Power analysis in mixed-effects modeling is frequently conducted by simulations also. For instance, Afshartous (1995) conducted a bootstrap analysis to determine a reasonable level-2 sample size that is required for stable and unbiased estimates of variance components in a two-level hierarchical linear model. Basically
he subsampled from over a thousand schools with a total of approximately 25,000 students contained in the data set for the National Educational Longitudinal Study in 1988. The analysis suggested that at least 320 groups were required to ensure correct variance-covariance estimates whereas only about 40 groups were needed to obtain reliable fixed effects estimates.

In an important study, Raudenbush and Liu (2000) investigated the relationship of sample sizes and statistical power for the random treatment effect in multisite randomized trials, also with a two-level hierarchical linear model. While holding other factors constant, statistical power was calculated as a function of treatment effect size with different numbers of sites as well as participants per site. The results indicated that increasing the level-2 sample size (i.e., sites) is more consequential on power than raising level-1 sample size (i.e., participants per site) especially when the variance of the treatment effect across sites is large. Note that in longitudinal studies, the level-1 units correspond to repeated measures within an individual while the level-2 units refer to individuals.

Based on the power calculation approach adopted by Raudenbush and Liu (2000), the Optimal Design software provides researchers with general guidelines for research design with regard to factors such as effect size, intra-class correlation, cluster size, and number of clusters, etc., in order to achieve satisfactory statistical power (see Spybrook, Raudenbush, Liu, & Congdon, 2006).
2.5 Other Important Simulation Studies

Other important simulation studies include investigation of the influence of the violation of assumptions on the mixed-effects modeling. The major assumption is, of course, the normality of the random effects, both at level-1 and level-2 (Raudenbush & Bryk, 2002; Verbeke & Molenberghs, 2000). It has been consistently pointed out that the mixed-effects model is quite robust against the violation of normality assumption (Zhang, 2005).
CHAPTER 3

METHODOLOGY

The first purpose of this study is to explore if the READ 180 program has a significant impact on reading proficiency of low-performing incarcerated youth. A longitudinal experimental design will be used for the evaluation of the intervention effect. The reading proficiency of the students in READ 180 and the comparison group will be measured using the Scholastic Reading Inventory (SRI). In addition, it is intended to check if the variability in the initial reading status as well as in rate of change of the target group could be explained by salient subject covariates such as age, gender, race/ethnicity, disability status, grade level, school mobility, and other assessments taken at the baseline including Reading and Math of the California Achievement Test (CAT). Mixed-effects modeling will be adopted for analysis, with three mean structures proposed for the reading growth pattern.

The other purpose of this study focuses on characterizing the sampling distribution of the estimate for the treatment effect of READ 180 on low-performing incarcerated youth, and methodologically guiding similar future longitudinal research with power analysis. Both the bootstrap technique and the Monte Carlo method will be adopted for this purpose, with the former based on the available sample and the
latter more oriented toward the target population. Power in this study is calculated both via an analytical approximation in SAS for the existing data and more importantly via simulations. Based on a series of Monte Carlo studies, power curves will be constructed to show how power differs when changing sample size, number of time points, Type I error rate, size of treatment effect, for both unbalanced and balanced data.

In this chapter, the population and sample are first described, followed by the instrumentation and data collection procedures. The statistical analysis plan and simulation designs are presented at the end.

3.1 Population and Sample

The target population of the study includes all low-performing incarcerated youth in a large mid-western state. The accessible population in this study includes all the 724 low-achieving incarcerated youth in the state who were determined to be eligible for the READ 180 program from the autumn of 2006 to the summer of 2007.

A judge in any juvenile court in this state can commit a youth to a public school district for a felony offense. The district is comprised of one intake facility and seven high schools, each located at one of the juvenile correctional facilities dispersed throughout the entire state. Among the seven schools, six are boys schools and one is girls school. These institutions operate year-round and offer four 10-week terms of schooling in each academic year. Each term has approximately 45 instructional days, and the school year begins in July and ends in June.
Starting from the autumn of 2006, the juvenile delinquents entered the school district though a process called intake during which they took the SRI, ReadCAT, and MathCAT. After assessment and record gathering, they were placed in a school and curriculum that was best suited for them. Based on the baseline SRI scores, they were determined for their eligibility for the READ 180 program. The accessible targeted population (N=724) included those who were found to be eligible. The demographic information for these youth is presented in Table 3.1. It can be seen that the absolute majority of this group were male and there was a high proportion of students having various disabilities. Information regarding race/ethnicity shows that there was a high percentage of Black and a considerable proportion of White youth, but very low percentages for other groups. The age ranged from 13 to 21, with most youth around 15 to 19 years old. As for the current grade levels, the majority of students were placed in grades 9 and 10. With regard to the special education status, it should be noted that in order to receive special education one must have a disability, whereas a youth with a disability does not have to be enrolled in special education. By examining the data, it was found that there were only two students with disability who did not receive special education services. Therefore, the information for special education (yes or no) can be regarded as a dichotomized variable for students’ disability status as the two are almost completely confounded.

The eligible youth were then randomly assigned to either the READ 180 group (experimental), or a “traditional” reading program (comparison). The randomization is verified by checking the baseline equivalence for the SRI and two
CAT scores. No difference was found between the two groups of youth for any of the three baseline tests. The youth remained in the public school district for different time lengths depending upon the nature of the offenses they committed and their behavior during incarceration. For various reasons, however, some eligible youth did not receive any or enough treatment although they had been assigned to one of the two groups. Therefore, the study sample was selected based on the criterion that the youth actually received at least two terms of instruction by either the READ 180 or the traditional program. A youth was considered to receive treatment in a given term if he actually attended at least one half of the instructional sessions in that term.

It is crucial to examine if the study sample was representative of the accessible population since random sampling was not used. The demographic information for the eligible incarcerated youth who received at least two terms of treatment and those with less than two terms of treatment is presented in Table 3.2 and Table 3.3 respectively. The similarity between the two groups in terms of these characteristics was checked using a chi-square test. It was found that the study sample was similar to the non-participating group in terms of race/ethnicity ($\chi^2_{(1)} = 2.09, p = .15$), gender ($\chi^2_{(1)} = .52, p = .47$), age ($\chi^2_{(8)} = 9.22, p = .32$), grade level ($\chi^2_{(5)} = 10.01, p = .08$), disability ($\chi^2_{(1)} = 1.78, p = .18$), and also treatment group ($\chi^2_{(1)} = .04, p = .84$). Based on a $t$-test, it was also found that the study participants were similar to the non-participants with regard to the baseline SRI ($t = .17, p = .86$), ReadCAT ($t = 1.35, p = .18$), and MathCAT ($t = .26, p = .79$). Therefore, the study sample can be regarded as representative of the accessible population.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>READ 180</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Freq</td>
<td>%</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>0</td>
<td>0.00%</td>
<td>1</td>
</tr>
<tr>
<td>Black</td>
<td>266</td>
<td>65.68%</td>
<td>217</td>
</tr>
<tr>
<td>Hispanic</td>
<td>12</td>
<td>2.96%</td>
<td>4</td>
</tr>
<tr>
<td>Native American/Alaskan</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
</tr>
<tr>
<td>Multiracial</td>
<td>16</td>
<td>3.95%</td>
<td>8</td>
</tr>
<tr>
<td>Not Available</td>
<td>5</td>
<td>1.23%</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total Non-White</strong></td>
<td></td>
<td>299</td>
<td>73.83%</td>
</tr>
<tr>
<td><strong>Total White</strong></td>
<td></td>
<td>106</td>
<td>26.17%</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>386</td>
<td>95.31%</td>
<td>307</td>
</tr>
<tr>
<td>Female</td>
<td>19</td>
<td>4.69%</td>
<td>12</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0.25%</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>1.98%</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>37</td>
<td>9.14%</td>
<td>27</td>
</tr>
<tr>
<td>16</td>
<td>64</td>
<td>15.80%</td>
<td>66</td>
</tr>
<tr>
<td>17</td>
<td>89</td>
<td>21.98%</td>
<td>75</td>
</tr>
<tr>
<td>18</td>
<td>110</td>
<td>27.16%</td>
<td>87</td>
</tr>
<tr>
<td>19</td>
<td>76</td>
<td>18.77%</td>
<td>51</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>3.46%</td>
<td>6</td>
</tr>
<tr>
<td>21</td>
<td>6</td>
<td>1.48%</td>
<td>2</td>
</tr>
<tr>
<td><strong>Current Grade</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-enrolled</td>
<td>4</td>
<td>0.99%</td>
<td>0</td>
</tr>
<tr>
<td>6-8</td>
<td>10</td>
<td>2.47%</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>205</td>
<td>50.62%</td>
<td>178</td>
</tr>
<tr>
<td>10</td>
<td>90</td>
<td>22.22%</td>
<td>53</td>
</tr>
<tr>
<td>11</td>
<td>32</td>
<td>7.90%</td>
<td>33</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>3.70%</td>
<td>17</td>
</tr>
<tr>
<td>13</td>
<td>49</td>
<td>12.10%</td>
<td>33</td>
</tr>
</tbody>
</table>

Note: *Pre-enrolled is designated as unknown grade but will be determined at a later date.

Table 3.1: Demographic Information of All Low-Achieving Incarcerated Youth in a Large Mid-Western State in the School Year 2006-2007.
Table 3.1 continued

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>READ 180</th>
<th></th>
<th>Traditional</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Freq</td>
<td>%</td>
<td>Freq</td>
<td>%</td>
</tr>
<tr>
<td>Disability Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autism</td>
<td></td>
<td>1</td>
<td>0.25%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Cognitive Disability-Mental Retardation</td>
<td></td>
<td>32</td>
<td>7.90%</td>
<td>37</td>
<td>11.60%</td>
</tr>
<tr>
<td>Deafness</td>
<td></td>
<td>1</td>
<td>0.25%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Emotional Disturbance</td>
<td></td>
<td>87</td>
<td>21.48%</td>
<td>59</td>
<td>18.50%</td>
</tr>
<tr>
<td>Mental Retardation</td>
<td></td>
<td>1</td>
<td>0.25%</td>
<td>1</td>
<td>0.31%</td>
</tr>
<tr>
<td>Not Applicable</td>
<td></td>
<td>211</td>
<td>52.10%</td>
<td>173</td>
<td>54.23%</td>
</tr>
<tr>
<td>Other Impairment-Minor</td>
<td></td>
<td>10</td>
<td>2.47%</td>
<td>8</td>
<td>2.51%</td>
</tr>
<tr>
<td>Other Impairment-Major</td>
<td></td>
<td>1</td>
<td>0.25%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Orthopedic Impairment</td>
<td></td>
<td>1</td>
<td>0.25%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Speech or Learning Disability</td>
<td></td>
<td>2</td>
<td>0.49%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Specific Learning Disability</td>
<td></td>
<td>56</td>
<td>13.83%</td>
<td>41</td>
<td>12.85%</td>
</tr>
<tr>
<td>Traumatic Brain Injury</td>
<td></td>
<td>1</td>
<td>0.25%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Visual Impairment</td>
<td></td>
<td>1</td>
<td>0.25%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Total Nondisabled</td>
<td></td>
<td>211</td>
<td>52.10%</td>
<td>173</td>
<td>54.23%</td>
</tr>
<tr>
<td>Total Disabled</td>
<td></td>
<td>194</td>
<td>47.90%</td>
<td>146</td>
<td>45.77%</td>
</tr>
<tr>
<td>Special Education</td>
<td>No</td>
<td>212</td>
<td>52.35%</td>
<td>174</td>
<td>54.55%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>193</td>
<td>47.65%</td>
<td>145</td>
<td>45.45%</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
<td>READ 180</td>
<td>Traditional</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------------</td>
<td>----------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Freq</td>
<td>%</td>
<td>Freq</td>
<td>%</td>
</tr>
<tr>
<td>Race</td>
<td>Native American/Alaskan</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>120</td>
<td>66.30%</td>
<td>89</td>
<td>61.38%</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>6</td>
<td>3.31%</td>
<td>1</td>
<td>0.69%</td>
</tr>
<tr>
<td></td>
<td>Native American/Alaskan</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Multiracial</td>
<td>9</td>
<td>4.97%</td>
<td>4</td>
<td>2.76%</td>
</tr>
<tr>
<td></td>
<td>Not Available</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Total Non-White</td>
<td>135</td>
<td>74.59%</td>
<td>94</td>
<td>64.83%</td>
</tr>
<tr>
<td></td>
<td>Total White</td>
<td>46</td>
<td>25.41%</td>
<td>51</td>
<td>35.17%</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>173</td>
<td>95.58%</td>
<td>141</td>
<td>97.24%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>8</td>
<td>4.42%</td>
<td>4</td>
<td>2.76%</td>
</tr>
<tr>
<td>Age</td>
<td>13</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>1</td>
<td>0.55%</td>
<td>1</td>
<td>0.69%</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>17</td>
<td>9.39%</td>
<td>13</td>
<td>8.97%</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>27</td>
<td>14.92%</td>
<td>25</td>
<td>17.24%</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>39</td>
<td>21.55%</td>
<td>36</td>
<td>24.83%</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>46</td>
<td>25.41%</td>
<td>42</td>
<td>28.97%</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>41</td>
<td>22.65%</td>
<td>24</td>
<td>16.55%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>7</td>
<td>3.87%</td>
<td>3</td>
<td>2.07%</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>3</td>
<td>1.66%</td>
<td>1</td>
<td>0.69%</td>
</tr>
<tr>
<td>Current Grade</td>
<td>Pre-enrolled*</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>6-8</td>
<td>1</td>
<td>0.55%</td>
<td>3</td>
<td>2.07%</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>83</td>
<td>45.86%</td>
<td>70</td>
<td>48.28%</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>50</td>
<td>27.62%</td>
<td>29</td>
<td>20.00%</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>19</td>
<td>10.50%</td>
<td>23</td>
<td>15.86%</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>9</td>
<td>4.97%</td>
<td>10</td>
<td>6.90%</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>19</td>
<td>10.50%</td>
<td>10</td>
<td>6.90%</td>
</tr>
</tbody>
</table>

Note: *Pre-enrolled is designated as unknown grade but will be determined at a later date.

Table 3.2: Demographic Information of the Low-Achieving Incarcerated Youth Who Received at Least Two Quarters of Treatment in a Large Mid-Western State in the School Year 2006-2007.
Table 3.2 continued

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>READ 180</th>
<th></th>
<th>Traditional</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Freq</td>
<td>%</td>
<td>Freq</td>
<td>%</td>
</tr>
<tr>
<td>Disability Status</td>
<td>Autism</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Cognitive Disability-Mental Retardation</td>
<td>17</td>
<td>9.39%</td>
<td>12</td>
<td>8.28%</td>
</tr>
<tr>
<td></td>
<td>Deafness</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Emotional Disturbance</td>
<td>40</td>
<td>22.10%</td>
<td>33</td>
<td>22.76%</td>
</tr>
<tr>
<td></td>
<td>Mental Retardation</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Not Applicable</td>
<td>89</td>
<td>49.17%</td>
<td>75</td>
<td>51.72%</td>
</tr>
<tr>
<td></td>
<td>Other Impairment-Minor</td>
<td>3</td>
<td>1.66%</td>
<td>2</td>
<td>1.38%</td>
</tr>
<tr>
<td></td>
<td>Other Impairment-Major</td>
<td>1</td>
<td>0.55%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Orthopedic Impairment</td>
<td>1</td>
<td>0.55%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Speech or Learning Disability</td>
<td>1</td>
<td>0.55%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Specific Learning Disability</td>
<td>28</td>
<td>15.47%</td>
<td>23</td>
<td>15.86%</td>
</tr>
<tr>
<td></td>
<td>Traumatic Brain Injury</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Visual Impairment</td>
<td>1</td>
<td>0.55%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Total Nondisabled</td>
<td>89</td>
<td>49.17%</td>
<td>75</td>
<td>51.72%</td>
</tr>
<tr>
<td></td>
<td>Total Disabled</td>
<td>92</td>
<td>50.83%</td>
<td>70</td>
<td>48.28%</td>
</tr>
<tr>
<td>Special Education</td>
<td>No</td>
<td>89</td>
<td>49.17%</td>
<td>75</td>
<td>51.72%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>92</td>
<td>50.83%</td>
<td>70</td>
<td>48.28%</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
<td>READ180</td>
<td>Traditional</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------------------</td>
<td>----------</td>
<td>-------------</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Freq</td>
<td>%</td>
<td>Freq</td>
<td>%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native American/Alaskan</td>
<td></td>
<td>0</td>
<td>0.00%</td>
<td>1</td>
<td>0.57%</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td></td>
<td>0</td>
<td>0.00%</td>
<td>1</td>
<td>0.57%</td>
</tr>
<tr>
<td>Hispanic</td>
<td></td>
<td>6</td>
<td>2.68%</td>
<td>3</td>
<td>1.72%</td>
</tr>
<tr>
<td>Black</td>
<td></td>
<td>146</td>
<td>65.18%</td>
<td>128</td>
<td>73.56%</td>
</tr>
<tr>
<td>Multiracial</td>
<td></td>
<td>7</td>
<td>3.13%</td>
<td>4</td>
<td>2.30%</td>
</tr>
<tr>
<td>Not Available</td>
<td></td>
<td>5</td>
<td>2.23%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Total Non-White</td>
<td></td>
<td>164</td>
<td>73.21%</td>
<td>136</td>
<td>78.16%</td>
</tr>
<tr>
<td>Total White</td>
<td></td>
<td>60</td>
<td>26.79%</td>
<td>38</td>
<td>21.84%</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>213</td>
<td>95.09%</td>
<td>166</td>
<td>95.40%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>11</td>
<td>4.91%</td>
<td>8</td>
<td>4.60%</td>
</tr>
<tr>
<td>Age</td>
<td>13</td>
<td>1</td>
<td>0.45%</td>
<td>1</td>
<td>0.57%</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>7</td>
<td>3.13%</td>
<td>3</td>
<td>1.72%</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>20</td>
<td>8.93%</td>
<td>14</td>
<td>8.05%</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>37</td>
<td>16.52%</td>
<td>41</td>
<td>23.56%</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>50</td>
<td>22.32%</td>
<td>39</td>
<td>22.41%</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>64</td>
<td>28.57%</td>
<td>45</td>
<td>25.86%</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>35</td>
<td>15.63%</td>
<td>27</td>
<td>15.52%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>7</td>
<td>3.13%</td>
<td>3</td>
<td>1.72%</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>3</td>
<td>1.34%</td>
<td>1</td>
<td>0.57%</td>
</tr>
<tr>
<td>Current Grade</td>
<td>Pre-enrolled*</td>
<td>4</td>
<td>1.79%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>6-8</td>
<td>9</td>
<td>4.02%</td>
<td>2</td>
<td>1.15%</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>122</td>
<td>54.46%</td>
<td>108</td>
<td>62.07%</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>40</td>
<td>17.86%</td>
<td>24</td>
<td>13.79%</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>13</td>
<td>5.80%</td>
<td>10</td>
<td>5.75%</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>6</td>
<td>2.68%</td>
<td>7</td>
<td>4.02%</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>30</td>
<td>13.39%</td>
<td>23</td>
<td>13.22%</td>
</tr>
</tbody>
</table>

Note: * Pre-enrolled is designated as unknown grade but will be determined at a later date.

Table 3.3: Demographic Information of the Low-Achieving Incarcerated Youth Who Received Less than Two Quarters of Treatment in a Large Mid-Western State in the School Year 2006-2007.

Continued
Table 3.3 continued

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>READ 180</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq</td>
<td>%</td>
<td>Freq</td>
</tr>
<tr>
<td>Autism</td>
<td>1</td>
<td>0.45%</td>
<td>0</td>
</tr>
<tr>
<td>Cognitive Disability-Mental Retardation</td>
<td>15</td>
<td>6.70%</td>
<td>25</td>
</tr>
<tr>
<td>Deafness</td>
<td>1</td>
<td>0.45%</td>
<td>0</td>
</tr>
<tr>
<td>Emotional Disturbance</td>
<td>47</td>
<td>20.98%</td>
<td>26</td>
</tr>
<tr>
<td>Mental Retardation</td>
<td>1</td>
<td>0.45%</td>
<td>1</td>
</tr>
<tr>
<td>Not Applicable</td>
<td>122</td>
<td>54.46%</td>
<td>98</td>
</tr>
<tr>
<td>Other Impairment-Minor</td>
<td>7</td>
<td>3.13%</td>
<td>6</td>
</tr>
<tr>
<td>Other Impairment-Major</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
</tr>
<tr>
<td>Orthopedic Impairment</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
</tr>
<tr>
<td>Speech or Learning Disability</td>
<td>1</td>
<td>0.45%</td>
<td>0</td>
</tr>
<tr>
<td>Specific Learning Disability</td>
<td>28</td>
<td>12.50%</td>
<td>18</td>
</tr>
<tr>
<td>Traumatic Brain Injury</td>
<td>1</td>
<td>0.45%</td>
<td>0</td>
</tr>
<tr>
<td>Visual Impairment</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
</tr>
<tr>
<td>Total Nondisabled</td>
<td>122</td>
<td>54.46%</td>
<td>98</td>
</tr>
<tr>
<td>Total Disabled</td>
<td>102</td>
<td>45.54%</td>
<td>76</td>
</tr>
<tr>
<td>Special Education</td>
<td>No</td>
<td>123</td>
<td>54.91%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>101</td>
<td>45.09%</td>
</tr>
</tbody>
</table>
3.2 Instrumentation

The major instrument used in this study is the SRI. It is an objective reading comprehension test that assesses students' reading levels and helps guide instruction according to students' needs by tracking students' reading growth over time and matching readers with appropriate texts (Scholastic, 2005a). Its scores in terms of lexile points serve as the outcome measure for this study. The lexile framework was developed by Meta-Metrics based on the Rasch Item Response Theory Model to place the difficulties of items and the abilities of readers on the logit scale (Scholastic, 1999; Stenner, Burdick, Sanford, & Burdick, 2006; Wright & Stone, 1979). “Lexile” refers to the unit of measurement for determining the difficulty of texts and the reading level of readers (Scholastic, 2001). A lexile point is “equivalent to $1/1000^{th}$ of the difference between the comprehensibility of basal primers (the midpoint of first grade text) and the comprehensibility of an electronic encyclopedia (the midpoint of workplace text)” (Scholastic, 2001, p.16).

The “Interactive” version of the SRI is a computer-adaptive assessment that was employed for assessing the study subjects (Scholastic, 1999, 2001). It employs a three-phase approach to assess a student’s reading level: “Start, Step, Stop” (Scholastic, 2001, p.3). After the student starts the test, the test items become more difficult or easier according to the student’s performance. The test stops when the computer has sufficient information about the student’s level of reading comprehension. In administering the SRI, the computer continually adapts the test based on the student’s responses to the questions. (Scholastic, 2001) The item bank
of the SRI currently has about 5,000 multiple choice items, and the validity and reliability of the SRI is advocated by Scholastic as being well-established (Scholastic, 1999, 2001, 2005a). However, without item level data, it is hard to evaluate the properties of this test instrument for low-achieving incarcerated youth.

3.3 Data Collection and Variables in the Study

For each participant, the SRI measure was taken at baseline, before the start of the assigned reading programs, and then repeatedly recorded at the end of each term during the school year. The SRI score serves as the outcome measure for the present study. The CAT assessments of reading and math were also administered to the subjects at the baseline, and they serve as two major covariates. The treatment group (1 = READ 180, 0 = Comparison) to which an eligible youth was assigned was recorded as the primary predictor of interest in the study.

The demographic information listed in Tables 3.1, 3.2, and 3.3 for the youth was also collected during the intake process. Given the fact that there were very low percentages of youth who belonged to race/ethnicity other than Black or White, the race/ethnicity variable is collapsed to two categories, namely White (=1) and non-White (=0). Also recall that special education was almost completely confounded with disability status, and thus only a dichotomized variable is employed in this study for disability status (0 = no disability, 1 = disability). The actual age and current grade level are another two student characteristics included in the study.
In addition, the institution at which each youth was placed was also recorded. It is known that the schools were confounded with gender with 6 boys schools and 1 girls school, and also that there were very few female youth in the accessible population. Thus the gender variable is not appropriate to be included in the study. Instead, the institution variable will be included to assess if there was any influence from the schools. Moreover, it was noted that some youth moved once or twice during the school year. Thus a school mobility variable is created based on if there was any movement between the schools (0= no mobility, 1= with mobility). The variable is dichotomized since there were quite few students with 2 moves.

3.4 Data Analysis

Before conducting formal analysis, the data set was cleaned by checking entry error and missing data. Among 326 subjects who received at least two terms of treatment, 23 had less than three measures of the SRI scores either due to record error or the subjects did not take the test for some particular reason that was unknown to the researcher. Therefore, these subjects were removed from the study. In addition, there were 31 subjects missing the baseline CAT scores. They were also removed from the study based on the consideration that bias is introduced too often when imputation methods are used for missing pre-test covariates. Without information for post-test CAT scores, one should be particularly cautious about imputing these baseline scores. No missing data were found for other covariates. As a result, the final data set that was used for analysis contained a total of 272 subjects.
The mixed-effects model is adopted in this study for investigating the program impact of the READ 180 on low-performing incarcerated youth due to its methodological advantages.

3.4.1 Applying the Mixed-Effects Model

In this section, a general mixed-effects model will be specified in matrix form first, and it is constructed based on a two-stage analysis approach (Verbeke & Molenberghs, 2000). Complete mixed-effects models applied to the data in this study will be presented in the following sections with three assumptions made about the overall trend across time. Consequently, three candidate mixed-effects models with different mean structures will be proposed later.

\( Y_i \) is the response variable vector containing the repeated SRI scores for the subject \( i \) among a total of \( n \) subjects.

\[
Y_i = \begin{bmatrix}
y_{i0} \\
y_{i1} \\
\vdots \\
y_{IN_i}
\end{bmatrix}_{N^*_i \times 1}
\]  

(3.1)

where \( N_i \) is the number of 10-week instructional terms that subject \( i \) attends, and \( N^*_i = N_i + 1 \). Thus \( y_{i0} \) is the baseline SRI measure of subject \( i \) prior to treatment.

In the first stage, a parametric growth curve is constructed as follows:

\[
Y_i = Z_i \beta_i + \varepsilon_i
\]  

(3.2)
where $\mathbf{Z}_i$ is a $(N_i^+ \times q)$ matrix of known covariates, $\mathbf{\beta}_i$ is a $q$-dimensional vector of unknown subject-specific coefficients, and $\mathbf{\varepsilon}_i = [\varepsilon_{i1} \ \varepsilon_{i2} \ \cdots \ \varepsilon_{iN_i}]'$ is a $N_i^+$-dimensional vector of residual components that follows a normal distribution with a zero mean vector and a variance structure of $\mathbf{\Sigma}_i$.

In the second stage, a multivariate regression model is built to explain the observed variability between the subjects with respect to their subject-specific regression coefficients $\mathbf{\beta}_i$ for the predictors of interest. In matrix format, it will be

$$
\mathbf{\beta}_i = \mathbf{K}_i \mathbf{\beta} + \mathbf{b}_i
$$

(3.3)

where $\mathbf{K}_i$ is $(q \times p)$ a matrix of known covariates, $\mathbf{\beta}$ is a $p$-dimensional vector of unknown regression parameters, and $\mathbf{b}_i$ represents random effects that are specific to the subject $i$ and are assumed to be independent and follow a $q$-dimensional multivariate normal distribution with zero mean vector and general unstructured covariance matrix $\mathbf{\Phi}$.

Upon the two-stage analysis, the mixed-effects model is constructed by substituting (3.3) into (3.2)

$$
\mathbf{Y}_i = \mathbf{X}_i \mathbf{\beta} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{\varepsilon}_i
$$

(3.4)

where $\mathbf{X}_i = \mathbf{Z}_i \mathbf{K}_i$ is a $(N_i^+ \times p)$ matrix of known covariates.

Building the mixed-effects model includes modeling both the mean structure for fixed effects and the covariance structure for residuals. In the following section, three mean structures are proposed for model selection.
3.4.2 Modeling the Mean Structure

One widely accepted approach for analyzing repeated measures is to capture the patterns of change in the mean response over time in terms of polynomial trends. The means are modeled as an explicit function of time via this approach which is known for the capability of dealing with highly unbalanced data (Fitzmaurice et al., 2004).

In the present study, three trends will be proposed for the mean structure, namely, the linear, quadratic, and cubic. For each trend, the full model will be built accordingly.

3.4.2.1 Linear Trend over Time

A straight line is the simplest possible trend for capturing changes in the mean response over time. In this model, the slope for time can be directly interpreted as a constant change in the mean response for one unit change in time. (Fitzmaurice et al., 2004)

In this model, a linear growth curve is built in the first stage:

\[ y_{ij} = \alpha_i + j \beta_i + \varepsilon_{ij}, \text{ for } i = 1,2,...,n \text{ and } j = 0,1,2,...N_i \]  

(3.5)

where \( y_{ij} \) is the SRI scores for student \( i \) at time point \( j \). A multivariate regression model is then built in the second stage based on pertinent predictors:

\[
\alpha_i = \alpha_{0i} + \alpha_1(\text{WHITE}_i) + \alpha_2(\text{AGE}_i) + \alpha_3(\text{MATHCAT}_i) + \alpha_4(\text{READCAT}_i) + \\
+ \alpha_5(\text{DISB}_i) + \alpha_6(\text{GRDLVL}_i) + \alpha_7(\text{INST}_i) + \alpha_8(\text{MOBL}_i) + b_{0i}
\]
The second stage analysis is useful to explain the observed variability between the subjects with respect to their subject-specific regression coefficients for different covariates.

In matrix notation, the components of this linear mixed model with $q = 2$ and $p = 19$ are shown below:

$$
\beta_i = \beta_0 + \beta_1(\text{WHITE}_i) + \beta_2(\text{AGE}_i) + \beta_3(\text{MATHCAT}_i) + \beta_4(\text{READCAT}_i) + \beta_5(\text{DISB}_i) + \beta_6(\text{GRDLVL}_i) + \beta_7(\text{INST}_i) + \beta_8(\text{MOBL}_i) + \beta_9(\text{TRTGRP}_i) + b_{ii}
$$

$$
\beta_i = \alpha_i^T \beta_{2x1}
$$

$$
\mathbf{K}_i = \begin{bmatrix}
1 & \text{WHITE} & \cdots & \text{MOBL} & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & \cdots & 0 & 1 & \text{WHITE} & \cdots & \text{MOBL} & \text{TRTGRP}
\end{bmatrix}_{2x19}
$$

$$
\mathbf{b}_i = \begin{bmatrix}
b_{0i} \\
b_{ii}
\end{bmatrix}_{2x1}
$$

### 3.4.2.2 Quadratic Trend over Time

When the rate of change in one period is more rapid than during others, a quadratic curve can be used to describe the growth over time (Cudeck, 2007; Verbeke & Molenberghs, 2000).
In this model, a quadratic growth curve is built in the first stage:

$$y_{ij} = \alpha_i + j\beta_i + j^2\gamma_i + \epsilon_{ij} \quad (3.11)$$

and the following multivariate regression model is built in the second stage again using pertinent predictors:

$$\alpha_i = \alpha_0 \ + \ \alpha_1 \text{WHITE}_i \ + \ \alpha_2 \text{AGE}_i \ + \ \alpha_3 \text{MATHCAT}_i \ + \ \alpha_4 \text{READCAT}_i \ + \ \alpha_5 \text{DISB}_i \ + \ \alpha_6 \text{GRDLVL}_i \ + \ \alpha_7 \text{INST}_i \ + \ \alpha_8 \text{MOBL}_i + b_{0i}$$

$$\beta_i = \beta_0 \ + \ \beta_1 \text{WHITE}_i \ + \ \beta_2 \text{AGE}_i \ + \ \beta_3 \text{MATHCAT}_i \ + \ \beta_4 \text{READCAT}_i \ + \ \beta_5 \text{DISB}_i \ + \ \beta_6 \text{GRDLVL}_i \ + \ \beta_7 \text{INST}_i \ + \ \beta_8 \text{MOBL}_i + \beta_9 \text{TRTGRP}_i + b_{1i}$$

$$\gamma_i = \gamma_0 + \gamma_1 \text{WHITE}_i + \gamma_2 \text{AGE}_i + \gamma_3 \text{MATHCAT}_i + \gamma_4 \text{READCAT}_i + \gamma_5 \text{DISB}_i + \gamma_6 \text{GRDLVL}_i + \gamma_7 \text{INST}_i + \gamma_8 \text{MOBL}_i + \gamma_9 \text{TRTGRP}_i + b_{2i}$$

In matrix notation, the components of this mixed-effects model with $$q = 3$$ and $$p = 29$$ are shown in the following:

$$Z_i = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & N_i & N_i^2 \end{bmatrix}_{(N_i+1) \times 3} \quad (3.12)$$

$$\begin{bmatrix} \alpha_i \\ \beta_i \\ \gamma_i \end{bmatrix}_{3 \times 1} \quad (3.13)$$

$$K = \begin{bmatrix} 1 \text{WHITE} & \ldots & \text{MOBL} & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 1 \text{WHITE} & \ldots & \text{MOBL} & \text{TRTGRP} & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 1 \text{WHITE} & \ldots & \text{MOBL} & \text{TRTGRP} \end{bmatrix}_{3 \times 29} \quad (3.14)$$

$$\begin{bmatrix} \alpha_0 & \ldots & \alpha_8 & \beta_0 & \ldots & \beta_9 & \gamma_0 & \ldots & \gamma_9 \end{bmatrix}_{29 \times 1}^T \quad (3.15)$$

and
3.4.2.3 Cubic Trend over Time

The same two steps are followed when assuming a more complex cubic trend for the mean response over time.

\[ y_{ij} = \alpha_i + j\beta_i + j^2\gamma_i + j^3\delta_i + \varepsilon_{ij} \]  \hspace{1cm} (3.17)

where

\[ \alpha_i = \alpha_0 + \alpha_1(\text{WHITE}_i) + \alpha_2(\text{AGE}_i) + \alpha_3(\text{MATHCAT}_i) + \alpha_4(\text{READCAT}_i) \]
\[ + \alpha_5(\text{DISB}_i) + \alpha_6(\text{GRDLVL}_i) + \alpha_7(\text{INST}_i) + \alpha_8(\text{MOBL}_i) + b_{0i} \]

\[ \beta_i = \beta_0 + \beta_1(\text{WHITE}_i) + \beta_2(\text{AGE}_i) + \beta_3(\text{MATHCAT}_i) + \beta_4(\text{READCAT}_i) \]
\[ + \beta_5(\text{DISB}_i) + \beta_6(\text{GRDLVL}_i) + \beta_7(\text{INST}_i) + \beta_8(\text{MOBL}_i) + \beta_9(\text{TRTGRP}_i) + b_{1i} \]

\[ \gamma_i = \gamma_0 + \gamma_1(\text{WHITE}_i) + \gamma_2(\text{AGE}_i) + \gamma_3(\text{MATHCAT}_i) + \gamma_4(\text{READCAT}_i) \]
\[ + \gamma_5(\text{DISB}_i) + \gamma_6(\text{GRDLVL}_i) + \gamma_7(\text{INST}_i) + \gamma_8(\text{MOBL}_i) + \gamma_9(\text{TRTGRP}_i) + b_{2i} \]

\[ \delta_i = \delta_0 + \delta_1(\text{WHITE}_i) + \delta_2(\text{AGE}_i) + \delta_3(\text{MATHCAT}_i) + \delta_4(\text{READCAT}_i) \]
\[ + \delta_5(\text{DISB}_i) + \delta_6(\text{GRDLVL}_i) + \delta_7(\text{INST}_i) + \delta_8(\text{MOBL}_i) + \delta_9(\text{TRTGRP}_i) + b_{3i} \]

In matrix notation, the components of the above mixed-effects model with \( q = 4 \) and \( p = 39 \) would be:

\[
Z_i = \begin{bmatrix}
1 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 \\
\vdots & \vdots & \vdots & \vdots \\
1 & N_i & N_i^2 & N_i^3 \\
\end{bmatrix}_{(N_i+4)\times4}
\]  \hspace{1cm} (3.18)
\[ \beta_i = \begin{bmatrix} \alpha_i \\ \beta_i \\ \gamma_i \\ \delta_i \end{bmatrix}_{4 \times 1} \]  

(3.19)

\[ \kappa_i = \begin{bmatrix} 1 & \text{WHITE} & \cdots & \text{MOBL} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \]  

(3.20)

\[ \beta = \begin{bmatrix} \alpha_0 & \cdots & \alpha_8 & \beta_0 & \cdots & \beta_9 & \gamma_0 & \cdots & \gamma_9 & \delta_0 & \cdots & \delta_9 \end{bmatrix}^T_{39 \times 1} \]  

(3.21)

and

\[ b_i = \begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \\ b_{3i} \end{bmatrix}_{4 \times 1} \]  

(3.22)

3.4.3 Modeling the Covariance Structure

The other aspect of mixed-effects model building involves the modeling of the covariance structure of residuals, \( \Sigma_i \). For the subjects in this study, the maximum number of repeated measures for the SRI is 5. The commonly used covariance structures for \( \Sigma_i \) can be employed for modeling, as shown in Table 3.4.
<table>
<thead>
<tr>
<th>Structure Name</th>
<th>Matrix Form</th>
</tr>
</thead>
</table>
| Simple                               | \[
\begin{bmatrix}
\sigma^2 \\
0 & \sigma^2 \\
0 & 0 & \sigma^2 \\
0 & 0 & 0 & \sigma^2 \\
0 & 0 & 0 & 0 & \sigma^2
\end{bmatrix}
\]                                       |
| Unstructured                         | \[
\begin{bmatrix}
\sigma_1^2 \\
\sigma_{12} & \sigma_2^2 \\
\sigma_{13} & \sigma_{23} & \sigma_3^2 \\
\sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \\
\sigma_{15} & \sigma_{25} & \sigma_{34} & \sigma_{35} & \sigma_5^2
\end{bmatrix}
\]                                       |
| Variance Components or Banded (1)    | \[
\begin{bmatrix}
\sigma_1^2 \\
0 & \sigma_2^2 \\
0 & 0 & \sigma_3^2 \\
0 & 0 & 0 & \sigma_4^2 \\
0 & 0 & 0 & 0 & \sigma_5^2
\end{bmatrix}
\]                                       |
| Banded (2)                           | \[
\begin{bmatrix}
\sigma_1^2 \\
\sigma_{12} & \sigma_2^2 \\
0 & \sigma_{23} & \sigma_3^2 \\
0 & 0 & \sigma_{34} & \sigma_4^2 \\
0 & 0 & 0 & \sigma_{45} & \sigma_5^2
\end{bmatrix}
\]                                       |
| Compound Symmetry                    | \[
\begin{bmatrix}
\sigma_1^2 + \sigma^2 \\
\sigma_1^2 & \sigma_1^2 + \sigma^2 \\
\sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma^2 \\
\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma^2 \\
\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma^2
\end{bmatrix}
\]                                       |
| Toeplitz                             | \[
\begin{bmatrix}
\sigma^2 \\
\sigma_{12} & \sigma^2 \\
\sigma_{13} & \sigma_{12} & \sigma^2 \\
\sigma_{14} & \sigma_{13} & \sigma_{12} & \sigma^2 \\
\sigma_{15} & \sigma_{14} & \sigma_{13} & \sigma_{12} & \sigma^2
\end{bmatrix}
\]                                       |
| First-Order Autoregressive AR (1)    | \[
\begin{bmatrix}
\sigma^2 \\
\rho \sigma^2 & \sigma^2 \\
\rho^2 \sigma^2 & \rho \sigma^2 & \sigma^2 \\
\rho^3 \sigma^2 & \rho^2 \sigma^2 & \rho \sigma^2 & \sigma^2 \\
\rho^4 \sigma^2 & \rho^3 \sigma^2 & \rho^2 \sigma^2 & \rho \sigma^2 & \sigma^2
\end{bmatrix}
\]                                       |

Table 3.4: Commonly Used Covariance Structures for Residuals
3.4.4 Model Selection Procedures

Two approaches can be used for model selection in mixed-effects models using SAS PROC MIXED (Ngo & Brand, 1997). The first method essentially identifies all the possible combinations of mean structures for fixed effects and the covariance structures for residuals applicable to the study. Hence the number of possible models for consideration tends to be quite large in some cases. Each of these models is then fitted in SAS using ML estimation. The best model is chosen based on the calculated fit indices (e.g., AIC and BIC) for all the models. The ML estimation approach is used here due to the fact that REML estimation only allows for comparison of models that differ in their random effects (Snijders & Bosker, 1999).

The second method is performed in two steps. In the first step, the full model with the most complex mean structure under consideration is fitted with all the possible covariance structures in SAS using REML estimation. REML is used in this step in order to focus on the covariance structures for the random effects. The best covariance structure for residuals is then selected based on the fit indices. In the second step, models with all possible mean structures are fitted with the selected covariance structure in SAS using ML. ML should be used here to allow for comparisons between models with different mean structures for the fixed effects. A best mean structure is then chosen based on the fit indices for these models.

It has been demonstrated that the two approaches normally give identical results (Ngo & Brand, 1997). Therefore, the latter two-step approach will be used in the model selection process since it is more straightforward and efficient.
3.4.5 Power Calculation for the Existing Data Set

As mentioned before, the power calculation for an empirical data set can be performed in SAS using the PROC MIXED procedure via an analytical approximation based on the $F$-distribution with a noncentrality parameter (Littell et al., 2006). This is carried out from a pilot study perspective, aiming to guide future studies (Lenth, 2001; Hoenig & Heisey, 2001).

3.4.6 Outlier Analysis

In order to accurately identify outlying cases, at least two pieces of information are needed in this study: one is the number of items a student answers during a test, and the other is the length of time that a student takes to complete the test. However, both are currently unavailable to the researcher. Therefore, a rule of thumb may be employed to detect outliers: any subject having an absolute difference of 600 points (about 3 standard deviations) for two adjacent repeated measures of the SRI will be considered as outliers. Based on this rough rule, a total of 10 outlying cases are identified.

Later the results for the final model will be presented for data both with and without outliers in order to examine the possible influence of these cases.
3.5 Bootstrap Analysis

3.5.1 The Bootstrap Algorithm

One fundamental objective of statistical inference goes beyond estimating one parameter or a set of parameters of interest based on an observed sample, to obtaining properties, such as accuracy, bias, and confidence intervals, of the estimate(s). The bootstrap procedure is quite useful in achieving the goal of characterizing the sampling distribution of parameter estimates (Efron, 1979, 1981, 1986; Efron & Stein, 1981; Efron & Tibshirani, 1986).

An observed sample $X = (x_1, x_2, \ldots, x_n) \rightarrow$ Empirical distribution $\hat{F}$

$X^*_1, X^*_1, \ldots, X^*_B$

$\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_B$

Bootstrap resampling

Figure 3.1: A Schematic Diagram of the Bootstrap in One-Sample Conditions (adapted from Figure 8.1 in Efron & Tibshirani, 1993, p.87)
The general bootstrap technique refers to a resampling procedure in which a large number of samples are randomly drawn with replacement from an empirical distribution of an observed data set in order to obtain an approximated sampling distribution of the parameter estimate of interest (Efron & Tibshirani, 1993). The basic idea is depicted in Figure 3.1. By evaluating all the bootstrap replications, statistical properties of the parameter estimates, such as standard error, quantiles, and confidence intervals, can be calculated.

Different from most traditional estimation techniques, the bootstrap does not make any distributional assumptions on the population from which the data are collected, and only requires that the observed sample is representative of the underlying population (Kelley, 2005). Thus the bootstrap method is particularly useful when the normality assumption is not necessarily satisfied. Statisticians note that bootstrap estimates are usually as accurate as or even more accurate than asymptotic normal estimates of sampling distributions (Manski, 1996). Today the bootstrap is applied in many fields to answer complicated problems that cannot be handled effectively by conventional statistical analyses (Efron & Tibshirani, 1986, 1993). In the area of educational research, this technique is also recommended as an alternative to traditional estimation procedures (Hutchinson, Morrison, & Felgate, 2000; Maas & Hox, 2005).

Among variants of the bootstrap procedures, the nonparametric cases bootstrap is considered as the most appropriate for the present study. Essentially, the nonparametric cases bootstrap draws complete cases (outcome variables with their
predictors) with replacement from the observed data set. In this study, repeated measures of the SRI and all the covariates for a given subject are regarded as an entire case. This way the dependence among observations measured at different time points can be preserved. In the context of multilevel modeling, it is pointed out that the nonparametric cases bootstrap yield more accurate estimates of both fixed and random parameters and better empirical coverage probabilities of confidence intervals (Van der Leeden, Busing, & Meijer, 1996, 1997).

As for the model that is fitted to each bootstrapped sample, it should be the final mixed-effects model that is chosen based on both scientific model selection criteria and the principle of parsimony. Fortunately, with advances in computing power, the computer intensive bootstrap approach can be carried out in a fully automated way.

3.5.2 The Bootstrap Procedures in the Current Study

In this study, the bootstrap procedures are implemented as follows:

1. Generate $B = 1000$ independent bootstrap samples $X^*_1, X^*_2, \ldots, X^*_{1000}$. Each bootstrap sample is obtained by randomly selecting $n$ subjects with replacement from the original data set. The size of these bootstrap samples, $n$, is the same as the size of the original sample, i.e., 272. The generated bootstrap samples are stored separately in 1000 files. This process can be automated by use of a MATLAB script (see Appendix A).
2. Independently fit the final selected mixed-effects model to each bootstrap sample, \( X^*_1, X^*_2, \ldots, X^*_1000 \), using a SAS macro (see Appendix B). Again, the results for all bootstrap samples are stored separately in 1000 output files.

3. Statistical properties including mean, standard error, confidence intervals, and power of the sampling distribution of the estimated treatment effect are then calculated based on all the estimates from the bootstrapped samples, \( \hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(1000) \). A PERL script (see Appendix C) is used to extract all the estimates for independent bootstrap samples from 1000 SAS output files.

As for the third step, a series of formulas are employed to obtain the statistical properties based on the bootstrap replications (Efron & Tibshirani, 1993). For an estimate of a fixed effect in the mixed-effects model, the bootstrap estimated mean is calculated by

\[
\hat{\theta}^*(.) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b) \tag{3.23}
\]

where \( B \) is the number of bootstrap samples and in this case equal to 1000.

The bootstrap estimated standard error is computed by

\[
\text{SE}_{\text{bootstrap}} (\hat{\theta}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(.)]^2} \tag{3.24}
\]

The bootstrap estimated cumulative density function is derived based on

\[
\hat{F}_{\text{bootstrap}} (\theta) = \frac{\# \{ \hat{\theta}^* \leq \theta \}}{B} \tag{3.25}
\]
The bootstrap estimated $\tau$th quantile is $\hat{F}_{\text{bootstrap}}^{-1}(\tau)$, and the $(1-\alpha)\%$ confidence interval is denoted by

$$\left[\hat{F}_{\text{bootstrap}}^{-1}(\alpha/2), \hat{F}_{\text{bootstrap}}^{-1}(1-\alpha/2)\right]$$

(3.26)

In addition, the bootstrap calculated power for a given Type I error rate, $\alpha$, is given by

$$\pi = \frac{\#\{p\text{-value} (\hat{\theta}^*) \leq \alpha\}}{B}$$

(3.27)

3.6 Monte Carlo Studies

Although the aforementioned bootstrap is quite straightforward to carry out, it obtains the sampling distribution of the parameter estimate of interest based on an original data set only. Monte Carlo studies, on the other hand, can be conducted to assess statistical properties of an estimate’s distribution for the target population under different possible scenarios. For instance, the Monte Carlo method can be employed in this study to simulate samples that have different treatment effects, sample sizes, and numbers of time points. It can be used to solve many problems that bootstrap cannot.

The term “Monte Carlo” first appeared in the 1940s, and originally referred to simulations that involved random walks (Murdoch, 2000). Recently, a Monte Carlo study refers to any computer experimentation that involves the simulation of random numbers (Martinez & Martinez, 2002). The major challenge in utilizing the Monte Carlo method is that the simulation process needs to be designed very carefully. It is
a more difficult task than implementing the nonparametric bootstrap resampling. A series of Monte Carlo studies will be conducted in this study, and the design of various scenarios will be described later in detail. Also, the mixed-effect model fitted to the simulated data will be the same final model that is used in the bootstrap approach. It is assumed that the selected model fits closely in the target population. Since all models are “false” to some degree, one should aim at a “useful” model that relies on a small number of variables in the interest of parsimony rather than a “correct” one (Box, 1976, 1979).

Basically for each Monte Carlo study, data simulated for model fitting consist of predictors and outcome scores. The covariates such as Age, Disability Status, the CAT scores, etc., can be simulated based on the information from all the 724 incarcerated youth who were eligible for the READ 180 program during 2006-2007. The generated subjects with different covariates values will then be randomly assigned to either the READ 180 or the comparison group with equal probability (an optimal 50-50 allocation for the treatment group predictor). Since the estimates of fixed and random parameters can be obtained from the previous bootstrap analysis, the outcome scores for each simulated subject can be generated accordingly.

In each Monte Carlo study of the series, \( R = 1000 \) rounds are performed to generate 1000 samples with a size of \( n \). During each round, the data simulation process is repeated \( n \) times to generate a sample of \( n \) subjects.
3.6.1 Simulating Data for Covariates

Based on the preliminary analysis, it seems that other than the READ 180 treatment group predictor, the following covariates may play a role in accounting for the variability in the initial reading status and the growth rate of the low-performing incarcerated youth: Age, Disability Status, Grade Level, ReadCAT, and MathCAT. Therefore, the correlations among these covariates are investigated first based on all the 724 eligible incarcerated youth.

3.6.1.1 Correlations among the Covariates

It is expected that the covariates may correlate with one another, so the data for the covariates cannot be simulated simply by assuming that they follow independent normal distributions. The correlations among different covariates and the number of measurements for the accessible population are depicted in Figure 3.2. It is assumed in the Monte Carlo simulations that these correlations hold in the underlying population. According to Figure 3.2, ReadCAT and MathCAT scores approximately follow normal distributions, while Age and Grade Level follow discrete distributions. Disability Status follows a Bernoulli distribution. Due to the fact that MathCAT and ReadCAT scores are correlated with Age, Disability Status, and Grade Level, subjects’ characteristics (Age, Disability Status, and Grade Level) will be generated first and then their MathCAT and ReadCAT scores will be simulated based on these characteristics.
Figure 3.2: Correlation Matrix Plot for Age, Disability, ReadCAT, MathCAT, Grade Level, and Number of Measurements Based on the Information of the Accessible Population (N=724)

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>ReadCAT*</th>
<th>Disability</th>
<th>MathCAT*</th>
<th>Grade Level</th>
<th>Number of Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReadCAT*</td>
<td>0.20</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disability</td>
<td>-0.08</td>
<td>-0.37</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MathCAT*</td>
<td>0.20</td>
<td>0.65</td>
<td>-0.39</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade Level</td>
<td>0.55</td>
<td>0.37</td>
<td>-0.08</td>
<td>0.36</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Number of Measurements</td>
<td>0.09</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.06</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: * The relevant correlations are based on pairwise deletion due to missing data.

Table 3.5: Correlations between Age, Disability, ReadCAT, MathCAT, Grade Level, and Number of Measurements Based on the Information of the Accessible Population (N=724)
3.6.1.2 Simulating Age, Disability Status, and Grade Level

The correlation matrix among different covariates is shown in Table 3.5. It is known that Disability Status has little correlation with Age and Grade Level. Therefore, Disability Status for each subject can be generated by an independent Bernoulli distribution with the parameter $p_{\text{Disab}}$, which is 0.4696 estimated from the 724 eligible youth, i.e., Disability $\sim$ Bernoulli (0.4696).

In addition, Age and Grade Level have a fairly high correlation with each other ($r = 0.55$). Counts for grade levels at different ages are listed in Table 3.6 based on which the joint probability distribution of Age and Grade Level can be obtained (see Table 3.7 and Figure 3.3). As a result, the Age and Grade Level covariates can be simulated from this empirical joint discrete distribution, i.e.,

$$\begin{bmatrix} \text{Age} \\ \text{Grade Level} \end{bmatrix} \sim \hat{F}(\text{Age, Grade Level}) \quad (3.28)$$

<table>
<thead>
<tr>
<th>Age</th>
<th>Counts</th>
<th>Grade Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.6: Counts for Grade Levels at Different Ages
<table>
<thead>
<tr>
<th>Age</th>
<th>Grade Level</th>
<th>Probability Mass Function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>0.0089</td>
<td>0.0060</td>
</tr>
<tr>
<td>15</td>
<td>0.0104</td>
<td>0.0805</td>
</tr>
<tr>
<td>16</td>
<td>0.0015</td>
<td>0.1682</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>0.1399</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0.1057</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>0.0491</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.0030</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.7: Joint Probability Mass Function for Age and Grade Level

Figure 3.3: Joint Distribution of Age and Grade Level
3.6.1.3 Simulating ReadCAT and MathCAT Scores

Based on Table 3.5 and Figure 3.2, it is known that the ReadCAT and MathCAT scores highly correlate with each other ($r = 0.65$). Therefore, it is reasonable to assume that the ReadCAT and MathCAT scores follow a bivariate normal distribution. Moreover, the two CAT scores also have fair correlations with Age, Disability Status, and Grade Level. Thus ideally all subjects should be grouped into many different groups based on their Age, Disability Status, and Grade Level, and each group will have their own bivariate normal distributions for the two CAT scores. This way a total of 96 groups have to be considered during the simulation process. Due to the substantial correlation between Age and Grade Level ($r = 0.55$), however, the actual implementation of the Monte Carlo simulations herein groups subjects only according to their Disability Status and Age. Therefore, 16 groups are generated with different Ages (14-21) and Disability Status (0 or 1).

Specifically, the mean vector and the covariance matrix of the ReadCAT and MathCAT scores for each student group are estimated from the eligible incarcerated youth following the two steps below:

1. Group the 724 subjects into 16 groups according to their Age and Disability Status.
2. For each student group, calculate the mean vector $\hat{\mu}_{\text{Age,Disb}} \begin{pmatrix} \text{ReadCAT} \\ \text{MathCAT} \end{pmatrix}$ and covariance matrix of the two CAT scores, $\hat{\Sigma}_{\text{Age,Disb}} \begin{pmatrix} \text{MathCAT} \\ \text{ReadCAT} \end{pmatrix}$.
In other words, the ReadCAT and MathCAT scores for a given subject will be simulated based on a bivariate normal distribution with the mean vector and covariance matrix of the group that the subject belongs to, i.e.,

\[
\begin{pmatrix}
\text{MathCAT} \\
\text{ReadCAT}
\end{pmatrix}_{\text{Age,Disb}} \sim N\left(\begin{pmatrix}
\mu_{\text{MathCAT}} \\
\mu_{\text{ReadCAT}}
\end{pmatrix}_{\text{Age,Disb}}, \Sigma\right)
\]

(3.29)

3.6.2 Generating Outcome Scores

After simulating data for the covariates, the outcome scores for a simulated subject at each time point can then be calculated based on the selected mixed-effects model from the values for Age, Disability Status, Grade Level, ReadCAT, MathCAT, as well as fixed effects, random effects, and the residuals of the model.

The fixed effects of the final model will employ the estimates from the bootstrap simulation. For a given subject with simulated values of Age, Disability Status, Grade Level, ReadCAT, and MathCAT, the corresponding random effects \((b_{0i}, b_{1i}, \text{and } b_{2i})\) can be simulated from a multivariate normal distributions with mean vector of \(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\) and a covariance matrix \(S_{b_i} = \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{pmatrix}\), which again come from the bootstrap estimates \(\begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{pmatrix}\) ~ \(N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{pmatrix}\right)\). The residuals are generated from a normal distribution with zero mean and a variance estimated from bootstrap study, i.e., \(\varepsilon_{ij} \sim N\left(0, \hat{\sigma}_{\varepsilon}^2\right)\).
Ideally in longitudinal research, a balanced design is followed as expected so that all subjects are measured at the same occasions with no missing data (Maas & Snijders, 2003). In reality, however, unbalanced data are almost always obtained even though a study adopts a balanced design in the first place. Simulation of balanced data for the outcome is fairly straightforward, whereas that of unbalanced data may pose a challenge. Both designs are detailed in the following two subsections.

3.6.2.1 Simulating Balanced Data for the Outcome

The procedures when generating the outcome for a subject with $t$ measurements are as follows:

1. Start from initial time point $m = 1$;
2. Calculate the outcome score at $m$;
3. If $m$ is smaller than $t$, raise $m$ by 1 and continue from step 2; otherwise stop.

Figure 3.4 displays a schematic diagram for the process in which balanced data are simulated so that all subjects have an equal number of measurements.
3.6.2.2 Simulating Unbalanced Data for the Outcome

The incarcerated youth remain at a correctional educational facility for different lengths of time depending upon the nature of offenses they commit and sometimes they do not take a test for various reasons, so it is unlikely to obtain balanced data in this context. As described previously, the actual data available in this study are unbalanced.

According to Table 3.5 and Figure 3.2, the number of measurements on the SRI for the eligible incarcerated youth in this study does not seem to correlate with Age, Disability Status, Grade Level, and the CAT scores. Therefore, it can be simulated independently. The observed distribution of the number of measurements
of the outcome is shown in Table 3.8. Thus the attrition rate of the subjects after receiving $x$ measurements is calculated by

$$p(x) = \frac{q(x)}{\sum_{m=x}^{5} q(m)}, x = 0,\ldots, 4$$

(3.30)

We define $p(0) = 0$ due to the fact that all subjects received at least 1 measurement, namely, the baseline. Table 3.9 lists the observed attrition rate after receiving $x$ measurements. Because there were 5 measurements at the most, the attrition rate after receiving 5 measurements was not observed. Based on the observed attrition rate (Table 3.9), however, the relationship between $p(x)$ and $x$ can be represented by the following linear regression with zero intercept:

$$p(x) = ax \text{ where } a = 0.1171$$

(3.31)

The $R^2$ is as high as 0.9938 (see Figure 3.5).

<table>
<thead>
<tr>
<th># of measurements, $x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># of subjects</td>
<td>73</td>
<td>130</td>
<td>175</td>
<td>136</td>
<td>158</td>
</tr>
<tr>
<td>Proportion, $q(x)$</td>
<td>0.1086</td>
<td>0.1935</td>
<td>0.2604</td>
<td>0.2024</td>
<td>0.2351</td>
</tr>
</tbody>
</table>

Table 3.8: Empirical Distribution of the Number of Measurements of the Outcome

<table>
<thead>
<tr>
<th>Number of Measurements, $x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attrition rate, $p(x)$</td>
<td>0</td>
<td>0.1086</td>
<td>0.2170</td>
<td>0.3731</td>
<td>0.4626</td>
</tr>
</tbody>
</table>

Table 3.9: Observed Attrition Rate after Receiving $x$ Measurements
This finding will be used to predict the drop-out pattern of the eligible incarcerated youth beyond 5 measurement occasions. To be specific, one can simulate the distribution of the number of measurements $x$ based on the following probability mass function:

$$f(x) = p(x) \prod_{m=0}^{x-1} [1 - p(m)] = ax \prod_{m=0}^{x-1} (1 - am), x = 1, 2, \cdots, \infty$$ \hspace{1cm} (3.32)

where $a = 0.1171$. The simulated distribution is shown in Figure 3.6. It can be seen that almost all subjects will drop out beyond 9 measurement occasions. Thus later
when one wants to examine statistical power by increasing the time points for the study, it is reasonable to only add a small number since very few subjects are likely to have more than 7 measurements of the SRI.

To simulate unbalanced data for the outcome, the following procedures are implemented when generating scores for a subject with at least 3 measurements and at most $t$ measurements.

1. Start from initial time point $m = 1$;
2. Calculate the outcome score at $m$;
3. If $m$ is larger than or equal to 3, generate a random number, $b$, based on a Bernoulli distribution with the parameter of the subject attrition rate at $m$, i.e., $p(m) = am$;
4. If $b = 1$, stop;
5. If $b = 0$ and $m$ is smaller than $t$, raise $m$ by 1 and continue from step 2.

Figure 3.7 displays a schematic diagram for the process in which unbalanced data are simulated so that subjects have different number of measurements of the SRI.
Figure 3.6: Simulated Distribution for Number of Measurements
Figure 3.7: A Schematic Diagram of Monte Carlo Simulation for Unbalanced Data
3.6.3 General Procedures for the Monte Carlo Studies

The overall iterative procedures for each Monte Carlo study are implemented in the following three stages:

1. Generate \( R = 1000 \) independent Monte Carlo simulated samples \( X_1, X_2, \ldots, X_{1000} \) by use of the design described above. These samples are stored separately in 1000 files. This process can be automated by use of a MATLAB script (see Appendix D).

2. Independently fit the desired mixed-effects model to each of the 1000 samples, \( X_1, X_2, \ldots, X_{1000} \), using a SAS macro (see Appendix B). The results for all the samples are also stored separately in 1000 output files.

3. Statistical properties including mean, standard error, confidence intervals, and power of the sampling distribution of the estimated treatment effect are then calculated based on all the estimates from the Monte Carlo simulated samples, \( \hat{\theta}(1), \hat{\theta}(2), \ldots, \hat{\theta}(1000) \). A PERL script (see Appendix C) is used to extract all the estimates from 1000 SAS output files.

In the third stage, the following formulas are used to calculate the statistical properties. For an estimate of a fixed effect in the mixed-effects model, the estimated mean for the 1000 Monte Carlo simulated samples is computed by

\[
\hat{\theta}(.) = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}(r)
\]  

(3.33)
where \( R \) is the number of simulated rounds in each study, i.e., 1000.

The estimated standard error is calculated by

\[
SE_{MCS}(\hat{\theta}) = \sqrt{\frac{1}{R-1} \sum_{r=1}^{R} [\hat{\theta}(r) - \hat{\theta}(\cdot)]^2}
\]  

(3.34)

The estimated cumulative density function is derived based on

\[
\hat{F}_{MCS}(\theta) = \frac{\#\{\hat{\theta} \leq \theta\}}{R}
\]  

(3.35)

The estimated \( \tau^{th} \) quantile is \( \hat{F}_{MCS}^{-1}(\tau) \), and the \((1-\alpha)\%\) confidence interval is denoted by

\[
[\hat{F}_{MCS}^{-1}(\alpha / 2), \hat{F}_{MCS}^{-1}(1 - \alpha / 2)]
\]  

(3.36)

3.6.4 Power Calculation via Monte Carlo Simulations

Using Monte Carlo simulations, statistical power can be calculated very straightforwardly based on the null and alternative distributions of the treatment effect.

3.6.4.1 Null Treatment Effect Distribution

Monte Carlo simulation can be performed for characterizing the distribution of the null treatment effect by specifying the regression weight for the treatment effect \( \gamma_T = 0 \) while keeping all the other parameters unchanged. The distribution of \( \gamma_T \) under \( H_0 \) can be determined by
\[
\hat{F}_{H_0}(\gamma_T) = \frac{\# \{ \hat{\gamma}_T \leq \gamma_T \}}{R} \quad (3.37)
\]

This simulated null distribution is useful when calculating the \( p \)-value and power for a specific treatment effect. For a given regression weight of the treatment effect, \( \gamma_{T,h} \), its \( p \)-value is computed by

\[
P(\gamma \geq \gamma_{T,h} \mid H_0) = 1 - \hat{F}_{H_0}(\gamma_{T,h}) \quad (3.38)
\]

3.6.4.2 Alternative Treatment Effect Distributions

Furthermore, Monte Carlo simulations can be carried out with assigned alternative treatment effects to investigate power of the statistical test on the treatment effect at different sizes. For a specified size of the alternative treatment effect, \( \gamma_{T,h} \), \( R \) samples will be generated and its distribution can be calculated by

\[
\hat{F}_{H_a}(\gamma_T) = \frac{\# \{ \hat{\gamma}_T \leq \gamma_T \}}{R} \quad (3.39)
\]

The bias of the estimates of the treatment effect is given by

\[
bias(\gamma_T) = \gamma_{T,h} - \frac{1}{R} \sum_{r=1}^{R} \hat{\gamma}_T(r) \quad (3.40)
\]

The power of detecting the alternative treatment effect at a given Type I error, \( \alpha \), can be computed by

\[
\pi = 1 - \hat{F}_{H_a}(\gamma_c) \quad \text{where} \quad \gamma_c = \hat{F}_{H_0}^{-1}(1 - \alpha) \quad (3.41)
\]

In addition, power curves for each specified treatment effect can be plotted at different Type I error rates for various combinations of design factors.
3.6.5 Varying Design Factors for Power Comparisons

Power comparisons in this study focus on influences of a few key factors. The range of the Type I error rate, $\alpha$, will be set to fall within the scale of (.00, .10). While holding other factors constant, statistical power of the test of the treatment effect at different Type I error rates will be examined with change in the levels of one of the following three factors respectively: (1) magnitude of the treatment effect, (2) sample size, and (3) number of time points.

As discussed previously, the Monte Carlo method can be used to simulate both balanced and unbalanced data. The current study will mainly rely on unbalanced data for power comparisons as this approximates the reality much more accurately. The balanced data will be only used to show the advantage in design for theoretical purposes.

3.6.5.1 Magnitude of Treatment Effect

The magnitude of the treatment effect is a crucial factor in impact studies. One common way to measure this factor is using Cohen’s effect size. However, since the target population in the study is low-performing incarcerated youth, a large effect size would be very unlikely. A more practical and evidence-based approach to guiding future research in this context is to use the quantile estimates of the treatment effect based on the bootstrap analysis. Specifically, the magnitude of the treatment effect will be set to the .025, .25, .75, and .975 quantiles of the bootstrap sampling distribution. Additionally, the treatment effect will also be set to the null and the
bootstrap estimated mean. The corresponding Cohen’s effect sizes at these magnitudes will also be calculated to provide a common measure of size of treatment effect.

3.6.5.2 Sample Size

For Monte Carlo simulations, the sample size will be set to 150, 300, and 600 respectively. It is anticipated that the results for the sample size of 300 should be comparable to those based on the bootstrap analysis of the original sample (n=272). On the one hand, it is of interest to check if an increase in sample size will substantially improve power. On the other hand, it is also meaningful to investigate if a smaller sample size will considerably affect power. An intuitive and efficient way to explore the possible range of sample sizes is to reduce 300 by one half for a smaller size (i.e., 150) and double the number for a larger size (i.e., 600). This exploration may provide similar future studies with a reasonable sample size.

3.6.5.3 Number of Time Points

For methodological purposes, it is assumed that the final model selected to fit the simulated data holds beyond the original 5 time points. It is of interest to expand the dimension in time and check if more time points will increase the power of detecting treatment effects. This is essentially equivalent to an increase in the size of level-1 units. As described in the design for simulating unbalanced data, there is a remarkable drop-out rate over time for the eligible incarcerated youth. Thus the
number of time points will be increased only by 1 and 2 respectively to compare with the original 5 points in time.
CHAPTER 4

RESULTS

This study focuses on the investigation of the intervention effect of the READ 180 program on reading level of low-achieving incarcerated youth in a large mid-western state, including relevant power analysis for longitudinal research. Based on the repeated SRI scores collected in an experimental study and also the simulated data, this study addresses the following questions:

- Does the READ 180 program have a significant impact on the reading level of low-achieving incarcerated youth over time, when comparing the experimental group with the comparison group being instructed with the traditional educational reading program on reading achievement and growth?
- Can the variability in the initial reading status of low-performing incarcerated youth be attributed to subject characteristics including age, gender, race/ethnicity, disability status, grade level, special education status, school mobility, and other baseline assessments including Reading and Math of the California Achievement Test (CAT)?
- Can the variability in the reading growth rate of low-performing incarcerated youth be accounted for by subject characteristics including age, gender,
race/ethnicity, disability status, grade level, special education status, school mobility, and other baseline assessments including Reading and Math of the CAT?

- What are the statistical properties of the probability distribution of the READ 180 intervention effect estimate, in terms of mean, standard error, and confidence intervals?

- What is power of the statistical test of the READ 180 intervention effect, based on both the available sample from a pilot study perspective and the simulations?

- Based on Monte Carlo studies, how does this power change in the context of longitudinal research when varying sample size, number of time points, Type I error rate, magnitude of treatment effect, and data pattern?

This chapter provides results with respect to the above research questions in four sections. Section 4.1 presents the exploratory analysis of the study sample. Section 4.2 details the formal modeling process of the longitudinal data, with comparisons between different covariance structures and growth curves. The final model is chosen based on both model selection criteria and the principle of parsimony. Section 4.3 describes the results based on the bootstrap analysis. Section 4.4 contains the results from a series of Monte Carlo studies for power comparisons under different circumstances.
4.1 Exploratory Analysis

Before any formal model building, exploratory analysis of the data is necessary in that it can provide useful summary descriptions.

Table 4.1 lists the mean SRI scores at different measurement occasions for all subjects and also for the two groups respectively. It can be seen from Figure 4.1 that there are apparent differences between the mean performances of the two groups over time. The READ 180 group on average steadily improves the SRI scores over time, while the mean performance of the comparison group increases at the beginning, especially after the first 10-week instructional period, but deceases substantially after the second term. It is also noted that after the first term, the comparison group actually outperforms the READ 180 group by 30.75 points, while the READ 180 group outperforms their counterparts by 40.19 points at the end of the fourth term.

The fact that the targeted experimental group lags behind during the first two terms is not terribly surprising when taking the length of instruction into account. It is believed that a twenty-week instructional period is needed to see noticeable improvement for struggling adolescent readers receiving the READ 180 program. Research done by the Scholastic (2006b) indicates that treatment exposure should encompass one academic year in order to observe marked reading improvement from the READ 180 program. Therefore, incarcerated youth who especially do not adapt well to change are anticipated to take an even longer time to adjust to a newly implemented program. Past data for READ 180 recipients suggest that the growth pattern shown in Figure 4.1 is expected. According to the descriptive statistics
presented in Table 4.1, it is known that after one year, the mean SRI score for the READ 180 group increases by 79.33 points whereas that for the comparison group only increases by 46.65, compared to their respective mean baseline scores.

Table 4.2 presents the standard deviations of the SRI scores at different measurement occasions for the two subgroups and the overall. It can be seen that the inter-individual variability of the outcome is quite large, as the standard deviations are consistently above 200 points beyond the baseline. It also seems that after the subjects receive either the READ 180 or the traditional English classes, the reading levels for these youth get a little more variable. In addition, the variance of the SRI scores appears to be relatively stable over time for both groups although the comparison group has a slightly larger variance. Hence it tells us that a constant variance structure could be a plausible starting point for modeling.

Moreover, the autocorrelations among the five repeated measures of the SRI are presented in both Figure 4.2 and Table 4.3. It can be observed that the correlations between two adjacent SRI scores are consistently stronger over time, as all these values are pretty large. The association between the baseline score and the SRI score measured after the first term is .70, a little weaker than other autoregressive weights of the first order. This is justifiable since the subjects are exposed to reading programs after being incarcerated. Correlations beyond one time point are still high but generally somewhat weaker than the first order autoregressive weights. One exception is the correlation between the SRI measured after the second term and that measured after one year, and the value is as large as .73.
Figure 4.1: Time Plot of the Mean Responses for the READ 180 Group, the Comparison Group, and the Overall

Table 4.1: Mean SRI Scores at Different Measurement Occasions for the READ 180 Group, the Comparison Group, and the Overall.

<table>
<thead>
<tr>
<th></th>
<th>SRI₀</th>
<th>SRI₁</th>
<th>SRI₂</th>
<th>SRI₃</th>
<th>SRI₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>READ 180</td>
<td>786.40</td>
<td>800.55</td>
<td>829.79</td>
<td>854.83</td>
<td>865.73</td>
</tr>
<tr>
<td>Comparison</td>
<td>778.89</td>
<td>831.30</td>
<td>837.78</td>
<td>819.95</td>
<td>825.54</td>
</tr>
<tr>
<td>Overall</td>
<td>783.28</td>
<td>813.36</td>
<td>833.09</td>
<td>841.70</td>
<td>851.75</td>
</tr>
</tbody>
</table>

Table 4.2: Standard Deviations of SRI Scores at Different Measurement Occasions for the READ 180 Group, the Comparison Group, and the Overall.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
<th>Term 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>READ 180</td>
<td>190.86</td>
<td>232.31</td>
<td>244.05</td>
<td>238.76</td>
<td>237.60</td>
</tr>
<tr>
<td>Comparison</td>
<td>205.53</td>
<td>248.46</td>
<td>258.59</td>
<td>289.15</td>
<td>272.01</td>
</tr>
<tr>
<td>Overall</td>
<td>196.75</td>
<td>239.19</td>
<td>249.72</td>
<td>258.80</td>
<td>249.85</td>
</tr>
</tbody>
</table>
Figure 4.2: Autocorrelation Scatter-Plot Matrix for Five Repeated Measures of the SRI for All Subjects.

Table 4.3: Autocorrelations for Five Repeated Measures of the SRI for All Subjects.

<table>
<thead>
<tr>
<th></th>
<th>SRI₀</th>
<th>SRI₁</th>
<th>SRI₂</th>
<th>SRI₃</th>
<th>SRI₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRI₀</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRI₁</td>
<td>0.70011</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRI₂</td>
<td>0.57687</td>
<td>0.73077</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRI₃</td>
<td>0.55526</td>
<td>0.62008</td>
<td>0.76784</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>SRI₄</td>
<td>0.57396</td>
<td>0.5965</td>
<td>0.7261</td>
<td>0.79364</td>
<td>1</td>
</tr>
</tbody>
</table>
After examining the mean pattern of the outcome scores, it is also of interest to look at intra-individual variability over time by investigating whether the response to treatment is homogeneous for each subject. The spaghetti plots shown in the left panels of Figures 4.3, 4.4, and 4.5 can provide useful information in this regard. It is noted that there is a striking range of individual differences in terms of reading growth. In other words, the individual records are heterogeneous, nearly idiosyncratic. The individual profiles are further divided into two groups of subjects who show a general increasing trend and those who display a general decreasing trend (see the middle and right panels of Figures 4.3, 4.4, and 4.5). It is observed that there are more subjects in the READ 180 group with overall positive slopes as compared to the comparison group. In addition, fewer subjects receiving the READ 180 program have overall negative slopes.

The exact numbers and percentages are also listed in Table 4.4. Around twenty-eight percent of the READ 180 students display a general negative trend, while almost forty percent in the comparison group have overall negative slopes. Based on these pieces of information, one might speculate that the READ 180 program may have a meaningful impact on the reading achievement of the target population.

Furthermore, according to the individual growth curves, the change in the SRI score at each occasion is unequal for most subjects, indicating a nonlinear trend. The spaghetti plots also suggest the presence of outlying cases in both the READ 180 group and the comparison group.
All the preliminary analyses suggest not only that there are dissimilar patterns in the mean performances of the two groups during the course of investigation, but also that there are apparent intra-individual variability over time.

Figure 4.3: Spaghetti Plots for **the Overall Group** in Total (Left Panel), Subjects with Positive Slopes (Middle Panel), and Subjects with Negative Slopes (Right Panel)
Figure 4.4: Spaghetti Plots for the READ 180 Group in Total (Left Panel), Subjects with Positive Slopes (Middle Panel), and Subjects with Negative Slopes (Right Panel).

Figure 4.5: Spaghetti Plots for the Comparison Group in Total (Left Panel), Subjects with Positive Slopes (Middle Panel), and Subjects with Negative Slopes (Right Panel).
<table>
<thead>
<tr>
<th></th>
<th>Overall n</th>
<th>col %</th>
<th>READ 180 n</th>
<th>col %</th>
<th>Comparison n</th>
<th>col %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope &gt; 0</td>
<td>183</td>
<td>67.28%</td>
<td>115</td>
<td>72.33%</td>
<td>68</td>
<td>60.18%</td>
</tr>
<tr>
<td>Slope ≤ 0</td>
<td>89</td>
<td>32.72%</td>
<td>44</td>
<td>27.67%</td>
<td>45</td>
<td>39.82%</td>
</tr>
<tr>
<td>Total</td>
<td>272</td>
<td>100%</td>
<td>159</td>
<td>100%</td>
<td>113</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 4.4: Number and Percentage of Subjects with Positive and Negative Growth Slopes in the READ 180 Group, the Comparison Group, and the Overall

When conducting exploratory analysis, it is also informative to briefly compare the adequacy of the three proposed first-stage models (i.e., linear, quadratic, and cubic trends) by using the coefficient of multiple determination $R^2$ to evaluate the overall goodness of fit for each parametric trend (Verbeke & Molenberghs, 2000).

The coefficient of multiple determination $R^2_i$ for subject $i$ is defined by

$$R^2_i = \frac{SSTO_i - SSE_i}{SSTO_i}$$

(4.1)

where

$$SSTO_i = (\mathbf{Y}_i - \mathbf{1}_i \mathbf{1}' / N_i)' (\mathbf{Y}_i - \mathbf{1}_i \mathbf{1}' / N_i)$$

$$SSE_i = (\mathbf{Y}_i - \mathbf{X}_i (\mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{X}_i' \mathbf{Y}_i)' (\mathbf{Y}_i - \mathbf{X}_i (\mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{X}_i' \mathbf{Y}_i)$$

and $\mathbf{1}_i = 1_{N_i \times 1}$.

It represents the proportion of the within-subject variability of the subject $i$ explained by the first-stage model and can be employed to measure the overall goodness of fit.

The overall goodness of fit for the model can be evaluated by the overall coefficient $R^2_{meta}$ of multiple determination, which can be obtained by
\[ R^2_{\text{meta}} = \frac{\sum_{i=1}^{n} (\text{SSTO}_i - \text{SSE}_i)}{\sum_{i=1}^{n} \text{SSTO}_i} \]  

\( R^2_{\text{meta}} \) represents the portion of the within-subject variability of all subjects explained by the model (Verbeke & Molenberghs, 2000). The larger \( R^2_{\text{meta}} \) is, the better the model describes the observed data. The information related to the subject-specific coefficients of multiple determination and the overall coefficient of multiple determination is plotted in Figure 4.6 for the linear, quadratic, and cubic trends respectively. It can be seen that when fitting a linear trend at the first-stage, the \( R^2_{\text{meta}} \) is only a little above .50, suggesting inadequacy. When a quadratic trend is used, this value is improved to almost .80. A cubic trend would produce an even larger \( R^2_{\text{meta}} \) of nearly .90.

Therefore, the rough measures based on preliminary analysis suggest that the growth pattern is curvilinear rather than linear. This useful information will be used during formal model selection.
Figure 4.6: Subject-Specific Coefficients $R^2_i$ of Multiple Determination and the Overall Coefficient $R^2_{meta}$ of Multiple Determination (Dashed Lines) for First-Stage Models which Assume Linear (Left Panel), Quadratic (Middle Panel), and Cubic (Right Panel) Subject-Specific Profiles
4.2 Results based on Mixed-Effects Modeling

Mixed-effects model building includes the modeling of both the mean structure for fixed effects and the covariance structure for random effects. The model selection procedures described in Section 3.4.4 will be employed to choose the final model. In the first step, the full model with the most complex mean structure under consideration is fitted with all the possible covariance structures for residuals in SAS using the REML estimation. The best covariance structure is then selected based on fit indices. In the second step, models with all possible mean structures are fitted with the selected covariance structure in SAS using the ML approach.

4.2.1 Selecting the Covariance Structure

When fitting the proposed most complex mean structure, namely the cubic trend full model, with all the seven commonly used covariance structures for the residuals, none of the covariance structures for the level-2 random effects $\Phi$ is positive definite. It seems that the cubic full model is overly complex and meanwhile there are not enough degrees of freedom. Even after removing the non-significant fixed effects, the cubic reduced model does not provide a positive definite $\Phi$ matrix either. Therefore, the quadratic trend full model is fitted with the seven proposed covariance structures using REML.

The corresponding fit indices are shown in Table 4.5. It is worth mentioning that among all the 7 possible models, only the two models with simple and banded (1)
structures respectively produce a positive definite $\Phi$ matrix. The model with Toeplitz covariance structure for the residuals does not even converge using REML.

It can be seen that the fit indices for the models with different covariance structures for the residuals are very close. This indicates that the covariance structure of the residuals does not affect the model fitting considerably in this case. Therefore, the simple covariance structure is chosen for the residuals.

This decision is made based on the following considerations: First, only simple and the banded (1) structures yield positive definite $\Phi$ matrices. Second, the models with the above two covariance structures produce very close AIC and BIC values and very similar results. All the conclusions that will be later drawn from the final model based on simple covariance structure are the same as those made based on banded (1) structure. Third, the model with simple covariance structure for the residuals is of less computational expense than the one with banded (1) covariance structure. The banded (1) structure takes about 4 times as much time as the simple one. Therefore, the bootstrap and Monte Carlo simulation studies are much more affordable when the simple covariance structure is used.

Put simply, the simple covariance structure fits the data sufficiently well, and will be adopted for the residuals. This structure will therefore be employed for choosing the best mean structure in the next step and also be used later for simulation studies.
Table 4.5: Fit Indices for Different Error Covariance Structures $\Sigma$ for Quadratic Models with Full Mean Structure of Fixed Effects Using REML Estimation.

<table>
<thead>
<tr>
<th>Structure</th>
<th># of Parameters</th>
<th>$-2$ (log-likelihood)</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>7</td>
<td>14914.8</td>
<td>14928.8</td>
<td>14954.1</td>
</tr>
<tr>
<td>AR (1)*</td>
<td>8</td>
<td>14896</td>
<td>14910</td>
<td>14935.3</td>
</tr>
<tr>
<td>Compound*</td>
<td>8</td>
<td>14914.8</td>
<td>14930.8</td>
<td>14959.7</td>
</tr>
<tr>
<td>Symmetry</td>
<td>8</td>
<td>14914.8</td>
<td>14930.8</td>
<td>14959.7</td>
</tr>
<tr>
<td>Toeplitz*</td>
<td>10</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Banded (1)*</td>
<td>11</td>
<td>14890.1</td>
<td>14912.1</td>
<td>14951.8</td>
</tr>
<tr>
<td>Banded (2)*</td>
<td>15</td>
<td>14881</td>
<td>14911</td>
<td>14965.1</td>
</tr>
<tr>
<td>Unstructured*</td>
<td>21</td>
<td>14881</td>
<td>14923</td>
<td>14998.7</td>
</tr>
</tbody>
</table>

Note: * The models with these four covariance structures for residuals do not produce a positive definite $\Phi$ matrix.

# The model with Toeplitz covariance structure for residuals does not converge.

4.2.2 Selecting the Mean Structure

The mean structure selection first involves the comparison of the full models for both linear and quadratic trend that are proposed in Section 3.4.2. The cubic trend model will not be considered even though its overall coefficient $R^2_{meta}$ of multiple determination approaches .90, because it does not yield a positive definite covariance matrix for the level-2 random effects.
4.2.2.1 Linear Growth Curve vs. Quadratic Growth Curve

When modeling the mean structure for the fixed effects, recall that ML estimation needs to be used during model fitting. The simple covariance structure for the residuals is applied based on the results from the first step analysis. The fit indices for the linear and quadratic models with full mean structure of fixed effects are shown in Table 4.6.

It can be seen that the AIC for the quadratic full model is substantially smaller than that for the linear full model, while the BIC for the former is a little larger than that of the latter. This is not too surprising because the AIC tends to favor a more complex model whereas the BIC prefers a more parsimonious one (Forster, 2000; McCoach & Black, 2008). A likelihood ratio test indicates that only a linear trend in time does not adequately account for the pattern of change over time, with an observed statistic of 59.7 and the $p$-value < .001. Also recall that the overall coefficient $R^2_{\text{meta}}$ of multiple determination measure for the linear trend is barely above .50, while the same measure for the quadratic trend improves remarkably to nearly .80. In addition, the change pattern displayed in the spaghetti plots shows that the trend is curvilinear rather than linear.

Based on the above information, the quadratic full model is preferred over the linear full model.

The statistically significant fixed effects in the two models are listed in Table 4.7 and Table 4.8 respectively. The covariance matrices for the random effects in these models are listed in Table 4.9. Efforts will not be devoted to interpreting the
parameter estimates here because they do not come from the selected final model. Still, one can see that with regard to the major parameter of interest, namely the treatment effect, is not found to be significant in the linear full model with a $p$-value of .08 (see Table 4.7). In contrast, the quadratic full model detects a significant interaction between the treatment predictor and the quadratic growth rate over time ($p$-value = .03, see Table 4.8).

<table>
<thead>
<tr>
<th>Model</th>
<th># of Parameters</th>
<th>-2 (log-likelihood)</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Full</td>
<td>33</td>
<td>15237.1</td>
<td>15303.1</td>
<td>15422.1</td>
</tr>
<tr>
<td>Quadratic Full</td>
<td>51</td>
<td>15177.4</td>
<td>15279.4</td>
<td>15463.3</td>
</tr>
</tbody>
</table>

Table 4.6: Fit Indices for Linear and Quadratic Models with Full Mean Structure of Fixed Effects and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate</th>
<th>SE</th>
<th>$t$-ratio</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>26.94</td>
<td>8.9824</td>
<td>3.00</td>
<td>0.0030</td>
</tr>
<tr>
<td>ReadCAT</td>
<td>30.58</td>
<td>6.0592</td>
<td>5.05</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Disability</td>
<td>-82.12</td>
<td>23.7807</td>
<td>3.45</td>
<td>0.0006</td>
</tr>
<tr>
<td>Disability*Time</td>
<td>20.1705</td>
<td>9.1692</td>
<td>2.20</td>
<td>0.0289</td>
</tr>
<tr>
<td>Treatment*Time</td>
<td>13.2974</td>
<td>7.5449</td>
<td>1.76</td>
<td>0.0794</td>
</tr>
</tbody>
</table>

Table 4.7: Estimates of Significant Fixed Effects for the Linear Model with Full Mean Structure and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation
<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>31.67</td>
<td>8.89</td>
<td>3.56</td>
<td>0.0004</td>
</tr>
<tr>
<td>ReadCAT</td>
<td>29.79</td>
<td>6.01</td>
<td>4.96</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Disability</td>
<td>-87.45</td>
<td>23.57</td>
<td>3.71</td>
<td>0.0003</td>
</tr>
<tr>
<td>MathCAT*Time</td>
<td>10.62</td>
<td>5.53</td>
<td>1.92</td>
<td>0.0561</td>
</tr>
<tr>
<td>Grade Level*Time²</td>
<td>-5.12</td>
<td>2.29</td>
<td>2.24</td>
<td>0.0264</td>
</tr>
<tr>
<td>Treatment*Time²</td>
<td>10.08</td>
<td>4.65</td>
<td>2.17</td>
<td>0.0314</td>
</tr>
</tbody>
</table>

Table 4.8: Estimates of Significant Fixed Effects for the Quadratic Model with Full Mean Structure and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Linear Full</th>
<th>Quadratic Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_0$</td>
<td>18123.00</td>
<td>15239.00</td>
</tr>
<tr>
<td>$b_1$</td>
<td>-314.52</td>
<td>2739.46</td>
</tr>
<tr>
<td>$b_2$</td>
<td>-14143.00</td>
<td>-1056.54</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td></td>
<td>12893.00</td>
</tr>
</tbody>
</table>

Note. * $p$-value < .05

Table 4.9: Estimated Covariance Matrices of the Random Effects for Linear and Quadratic Models with Full Mean Structure of Fixed Effects and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation
4.2.2.2 Quadratic Full Model vs. Quadratic Reduced Model

Even though the full model of quadratic trend is selected over that for the linear trend, it is noticed that there are too many non-significant fixed effects in the model with p-values much bigger than .05. Given this fact, the complex full model of quadratic trend is not straightforward enough for interpretation, and there is a need to select a more parsimonious final model.

Consequently, this full model is reduced by removing the non-significant fixed effects, leading to a more parsimonious quadratic model as follows:

\[
y_{ij} = \alpha_i + j\beta_i + j^2\gamma_i + \varepsilon_{ij}
\]  

(4.3)

where \(y_{ij}\) is the measurements of student \(i\) at time point \(j\) and

\[
\alpha_i = \alpha_1(\text{AGE}_i) + \alpha_2(\text{READCAT}_i) + \alpha_3(\text{DISB}_i) + b_{0i}
\]

\[
\beta_i = \beta_1(\text{MATHCAT}_i) + b_{1i}
\]

\[
\gamma_i = \gamma_1(\text{GRDLVL}_i) + \gamma_2(\text{TRTGRP}_i) + b_{2i}
\]

It is of great interest to see if this model provides better fit than the full model. Thus the reduced quadratic model is also fitted with a simple covariance structure for the residuals in SAS using ML. The corresponding fit indices are presented in Table 4.10 for comparison with the full model.

It can be seen that both AIC and BIC for the reduced model are smaller than those for the full. The AIC is only slightly smaller, largely because it is susceptible to picking the more complex model. The BIC, on the other hand, is improved a lot for the reduced model as expected. Based on these fit indices and more importantly the
principle of parsimony, it is decided that the quadratic reduced model is more favorable in this study.

The ML estimates for the fixed effects and the covariance matrix of the random effects of this model are listed in Table 4.11 and Table 4.12 respectively. According to Table 4.11, the significant fixed effects appeared in the quadratic full model are still statistically significant in the reduced model. When comparing the estimates of these parameters given by the reduced model to those from the full model, it can be observed that there are substantial differences in magnitudes. The Age covariate at the intercept term shows the smallest difference between the two estimates (31.67 vs. 33.45), while the other five fixed effects have much larger differences. The treatment effect of the READ 180 as the parameter of primary interest has a ML estimate of 6.04 for its interaction with the quadratic trend in the reduced model, whereas the sheer quantity of this parameter estimate is 10.08 in the full model. When looking at the standard errors for all the fixed effects estimates in the two models, it is obvious that the reduced model yields ML estimates with much smaller standard errors when compared to those given by the full model. The $p$-values for all the six significant fixed effects are also smaller in the reduced model than those in the full. Therefore, one has more confidence in the precision of the parameter estimates provided by the reduced model of quadratic trend.

All the above evidence indicates that the quadratic reduced model should be selected as the final model for interpretation in this study. The parsimony of this model can also benefit the simulation studies greatly.
Table 4.10: Fit Indices for Two Nested Quadratic Models with Full Mean Structure and Reduced Mean Structure of Fixed Effects and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation.

<table>
<thead>
<tr>
<th>Model</th>
<th># of Parameters</th>
<th>-2 (log-likelihood)</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>51</td>
<td>15177.4</td>
<td>15279.4</td>
<td>15463.3</td>
</tr>
<tr>
<td>Reduced</td>
<td>13</td>
<td>15251.8</td>
<td>15277.8</td>
<td>15324.6</td>
</tr>
</tbody>
</table>

Table 4.11: Estimates of Significant Fixed Effects for Two Nested Quadratic Models with Full Mean Structure and Reduced Mean Structure of Fixed Effects and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation.

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Quadratic Full</th>
<th>Quadratic Reduced</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>Age</td>
<td>$\alpha_1$</td>
<td>31.67</td>
</tr>
<tr>
<td>ReadCAT</td>
<td>$\alpha_2$</td>
<td>29.79</td>
</tr>
<tr>
<td>Disability</td>
<td>$\alpha_3$</td>
<td>-87.45</td>
</tr>
<tr>
<td>MathCAT</td>
<td>$\beta_1$</td>
<td>10.62</td>
</tr>
<tr>
<td>Grade Level</td>
<td>$\gamma_1$</td>
<td>-5.12</td>
</tr>
<tr>
<td>Treatment</td>
<td>$\gamma_2$</td>
<td>10.08</td>
</tr>
</tbody>
</table>

Table 4.12: Estimated Covariance Matrices of the Random Effects for Two Nested Quadratic Models with Full Mean Structure and Reduced Mean Structure of Fixed Effects and a Simple Error Covariance Structure $\Sigma$ Using ML Estimation.

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Quadratic Full</th>
<th>Quadratic Reduced</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$b_0$</td>
<td>$b_1$</td>
</tr>
<tr>
<td>$b_0$</td>
<td>15239.00*</td>
<td></td>
</tr>
<tr>
<td>$b_1$</td>
<td>2739.46</td>
<td>5045.63*</td>
</tr>
<tr>
<td>$b_2$</td>
<td>-1056.54*</td>
<td>-560.65</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>12893.00</td>
<td></td>
</tr>
</tbody>
</table>

Note. * $p$-value < .05
4.2.3 The Selected Final Model with Restricted Maximum Likelihood Estimation

The two-step model selection procedures identify the reduced model in (4.3) with the simple covariance structure for the residuals as the best model. Therefore, this model will be the main focus in the rest of the study.

A frequently mentioned drawback of ML estimation is that the estimates of variances and covariances are contingent on point estimates of the fixed effects (Raudenbush & Bryk, 2002), and most software sets REML as the default estimation method. Thus it is desirable to fit the above selected final model using REML for comparison with estimates given by ML. Relevant results are presented in Tables 4.13, 4.14, and 4.15 respectively.

It seems that the REML estimation provides relatively smaller fit indices for the final model (see Table 4.13). The corresponding REML estimates for the significant fixed effects in the final model are listed in Table 4.14. The results show that both ML and REML approaches lead to the same set of significant fixed effects (cf. Table 4.11), and that there is almost no difference in all the estimates and the corresponding standard errors provided by the two estimation methods. According to the literature, the two approaches typically yield similar results for fixed effects when the sample size is not particularly small (McCoach & Black, 2008). The REML estimates of variances and covariances for the random effects in Table 4.15 are expected to be more accurate though than the ML estimates listed in Table 4.12.

In general, the conclusions that can be drawn from the results based on the final model are essentially same whether the ML or the REML is used. Here the
REML estimates are used for interpretation as REML serves as the default estimation method used in most software.

Note that based on all the models that are compared, no significant results are found for the following three covariates: institution, school mobility, and race/ethnicity. This implies that none of these covariates can account for the variability in the initial reading status or the rate of reading growth, either in linear or quadratic trend, of the low-achieving incarcerated youth.

Based on the parsimonious quadratic reduced model, the interpretation of the fixed effects is fairly straightforward. First of all, with regard to the primary treatment predictor, the quadratic trend term is statistically significant ($\hat{\gamma}_2 = 6.06$, $t=3.00, p = .0031$). This result implies that the READ 180 program has a significantly positive impact on the reading proficiency of low-performing incarcerated youth. Moreover, it suggests that for the READ 180 program recipients, the rate of reading growth in one period is more rapid than during others instead of being constant. Based on the model fitting, the rate of change is quantified to be 6.06, and the actual growth after a certain amount of time needs to be multiplied by the squared time, i.e., SRI score gain = $6.06 \times \text{time}^2$. Specifically, this indicates that compared to subjects in the comparison group, the students exposed to the READ 180 program on average gain 6.06 more lexile points after the first term, 24.24 more points after the second term, 54.54 after the third, and 96.96 after one year, while controlling for other factors.
The fitted difference in the average SRI scores between the READ 180 and the comparison group is plotted in Figure 4.7. It can be seen that students in the READ 180 accelerate their reading growth as time goes by, compared to their counterparts. The same pattern regarding different reading growth over time for the two groups is captured more clearly in Figure 4.8. It basically shows that the reading gap between the two groups is widened over time.

Next the influence on the reading level of low-performing incarcerated youth from other meaningful covariates is examined. With regard to the initial reading status, the results indicate that Age, ReadCAT, and Disability Status can account for its variability among subjects. To be specific, on average a subject’s initial reading score is expected to increase 33.46 lexile points with one year’s increase in age ($\alpha_1 = 33.46, t=19.32, p <.0001$), and 38.27 points with one unit increase in the ReadCAT score ($\alpha_2 = 38.27, t = 9.19, p <.0001$). In addition, the average difference in initial reading between low-achieving incarcerated youth with disability and those without disability is 51.87 ($\alpha_3 = -51.87, t = -2.67, p =.008$).

Furthermore, MathCAT has a significantly positive effect on the rate of reading growth with a linear trend ($\beta_i = 5.22, t = 4.34, p <.0001$). That is, one unit increase in the MathCAT score is expected to constantly improve an average of 5.22 lexile points over each time period.

Grade Level, on the other hand, shows a significantly negative impact on the reading growth rate, displaying a quadratic trend ($\gamma_1 = -.78, t = -3.39, p =.0008$). This
indicates that the higher the grade level, the slower the growth. This pattern has been verified by Meta-Metrics. Students in higher grade levels have lower expected growth within a year. In this study, with one level’s increase in current grade, a low-performing incarcerated youth is expected to decelerate the reading growth by .78 lexiles after the first term, 3.12 after the second, 7.02 after the third, and 12.48 after one year on average, while controlling for other factors.

To observe the influence of these covariates more clearly, the mean functions for students with different characteristics (i.e., at different levels of these significant covariates) are plotted for two groups (READ 180 vs. Comparison) while keeping the parameters at their estimates (see Figures 4.9, 4.10, 4.11, 4.12, and 4.13).

In sum, the results based on fixed effects suggest that the READ 180 program appears to be qualitatively different from the traditional reading program for the comparison group. Moreover, the inference for other covariates makes sense and provides useful information regarding the variability in subjects’ initial reading status as well as rate of change over time.
<table>
<thead>
<tr>
<th>Model</th>
<th>-2 (log-likelihood)</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final (Quadratic Reduced)</td>
<td>15236.2</td>
<td>15250.2</td>
<td>15275.4</td>
</tr>
</tbody>
</table>

Table 4.13: Fit Indices for the Final Model Using REML Estimation

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age $\alpha_1$</td>
<td>33.46</td>
<td>1.73</td>
<td>19.32</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>ReadCAT $\alpha_2$</td>
<td>38.27</td>
<td>4.17</td>
<td>9.19</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Disability $\alpha_3$</td>
<td>-51.87</td>
<td>19.40</td>
<td>2.67</td>
<td>0.008</td>
</tr>
<tr>
<td>MathCAT $\beta_1$</td>
<td>5.22</td>
<td>1.20</td>
<td>4.34</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Grade Level $\gamma_1$</td>
<td>-0.78</td>
<td>0.23</td>
<td>3.39</td>
<td>0.0008</td>
</tr>
<tr>
<td>Treatment $\gamma_2$</td>
<td>6.06</td>
<td>2.02</td>
<td>3.00</td>
<td>0.0031</td>
</tr>
</tbody>
</table>

Table 4.14: Estimates for the Fixed Effects in the Final Model Using REML Estimation

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_0$</td>
<td>16968.00*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_1$</td>
<td>1732.72</td>
<td>7601.87*</td>
<td></td>
</tr>
<tr>
<td>$b_2$</td>
<td>-847.72</td>
<td>-1171.69</td>
<td>264.15</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td></td>
<td></td>
<td>12788.00*</td>
</tr>
</tbody>
</table>

Note. * $p$-value < .05

Table 4.15: Estimated Covariance Matrix for the Random Effects in the Final Model Using REML Estimation
Figure 4.7: Time Plot of the Fitted Mean Difference in the SRI Scores between the READ 180 Group and the Comparison Group for the Final Quadratic Reduced Model Using REML Estimation
Figure 4.8: Mean Functions for the READ 180 and the Comparison

Figure 4.9: Mean Functions for the READ 180 and the Comparison, at the .25 and .75 Quantiles of Age
Figure 4.10: Mean Functions for the READ 180 and the Comparison, with Disability and without Disability

Figure 4.11: Mean Functions for the READ 180 and the Comparison, at the .25 and .75 Quantiles of ReadCAT
Figure 4.12: Mean Functions for the READ 180 and the Comparison, at the .25 and .75 Quantiles of MathCAT

Figure 4.13: Mean Functions for the READ 180 and the Comparison, at the .25 and .75 Quantiles of Grade Level
So far one may be pretty satisfied with the information about the mean response provided by the quadratic reduced model. It would also be of interest to investigate how the final model works for different individuals, recalling the striking individual differences displayed in the spaghetti plots.

Figure 4.14 and Figure 4.15 present data and individual fitted quadratic regression equations for 25 subjects with the best model fit and another 25 with the worst model fit respectively. Model fit for individuals are evaluated based on the residual sum of squares.

It could be observed that most of the individuals with the worst model fit display either a flat slope or a dramatic bounce-up-and-down pattern. The latter pattern seems to come from the outlying cases in the study. Thus the next section explores the influence of the outliers on the results from the final model.
Figure 4.14: Trellis Plot for 25 Participants with the Best Model Fit, Each Graph with Data and Individual Fitted Quadratic Regression Equation
Figure 4.15: Trellis Plot for 25 Participants with the Worst Model Fit, Each Graph with Data and Individual Fitted Quadratic Regression Equation
4.2.4 Results for the Final Model after Removing the Outliers

Recall that the spaghetti plots support the existence of outlying cases in this study. Based on the rule of thumb discussed in Section 3.4.6, ten outliers were identified. One can now proceed to examine the influence of these extreme cases on the model fit and estimates of fixed effects.

The final model (quadratic reduced) is fitted again using REML approach after removing the 10 outliers. Relevant results are presented in Tables 4.16, 4.17, and 4.18 respectively. It is observed that there is a substantial decrease in fit indices after the outliers are deleted (see Table 4.16). This means that the model fits much better without the influence of these outliers.

Moreover, all the conclusions drawn previously based on the complete data set do not change after removing the outliers. According to the estimates for the fixed effects, they do not differ substantially from those in Table 4.14. The READ 180 treatment effect is now 5.88 and still statistically significant for the quadratic trend. A difference of .18 for this weight will not result in a big difference at all.

As for other covariates, most of them have very similar estimates compared to those obtained based on the complete data. The only noticeable difference is found for the estimate of influence of Disability Status on the initial reading performance. After removing the outliers, the average negative effect of Disability on student initial reading scores is decreased by 5.23 points (-46.64 vs. a previous -51.87).
Table 4.16: Fit Indices for the Final Model with REML Estimation after Removing Outliers.

<table>
<thead>
<tr>
<th>Model</th>
<th>-2 (log-likelihood)</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final (Quadratic Reduced) Outliers Removed</td>
<td>14398.9</td>
<td>14412.9</td>
<td>14437.9</td>
</tr>
</tbody>
</table>

Table 4.17: Estimates for the Fixed Effects in the Final Model with REML Estimation after Removing Outliers.

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate</th>
<th>SE</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age $\alpha_1$</td>
<td>33.88</td>
<td>1.74</td>
<td>19.52</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>ReadCAT $\alpha_2$</td>
<td>36.95</td>
<td>4.19</td>
<td>8.82</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Disability $\alpha_3$</td>
<td>-46.64</td>
<td>19.56</td>
<td>2.38</td>
<td>0.0178</td>
</tr>
<tr>
<td>MathCAT $\beta_1$</td>
<td>5.50</td>
<td>1.18</td>
<td>4.67</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Grade Level $\gamma_1$</td>
<td>-0.79</td>
<td>0.22</td>
<td>3.62</td>
<td>0.0004</td>
</tr>
<tr>
<td>Treatment $\gamma_2$</td>
<td>5.88</td>
<td>1.90</td>
<td>3.09</td>
<td>0.0023</td>
</tr>
</tbody>
</table>

Table 4.18: Estimated Covariance Matrix for the Random Effects in the Final Model with REML Estimation after Removing Outliers.

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_0$</td>
<td>20407.00*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_1$</td>
<td></td>
<td>-1066.50</td>
<td>12498.00*</td>
</tr>
<tr>
<td>$b_2$</td>
<td>-299.89</td>
<td>-2347.23*</td>
<td>550.44*</td>
</tr>
</tbody>
</table>

Note. * $p$-value < .05

Table 4.18: Estimated Covariance Matrix for the Random Effects in the Final Model with REML Estimation after Removing Outliers
4.2.5 Normality Check

The major assumption made by mixed-effects models is the normality for the random effects at both levels (Verbeke & Molenberghs, 2000). In other words, the level-1 residuals should have a univariate normal distribution, and the level-2 random effects should have a multivariate normal distribution. The multivariate normality requires not only that all univariate distributions are normal but also that any linear combinations of the variables are normally distributed.

The univariate normality assumption can be examined by looking at the normal Q-Q plots for each random effect in Figure 4.16. It is noted that the distribution of the level-1 residuals $\varepsilon$ is definitely not normal, but with heavy tails. The distributions of $b_0$ and $b_1$ have some departure from the reference line at the lower end, while the one of $b_2$ seems to be approximately normal.

A multivariate normality test is also carried out for $b_i$, as shown in Figure 4.17. For a sample size of 272 and 3 parameters, the test statistic is 41.62, with the $p$-value < .001. Therefore, the multivariate normality assumption does not hold for the level-2 random effects.

For a complex realistic data set like this one, the normality assumption is frequently violated. Fortunately, however, many studies consistently supports that the mixed-effects model is quite robust to the violation of the normality assumption (Zhang, 2005). In addition, the simulation studies will verify if this is the case in this study. As a result, the unmet assumption does not cause great concern at this time.
Figure 4.16: Side-by-Side Normality Q-Q Plots of the Random Effects

Figure 4.17: Multivariate Normality Test for $b_1$
4.3 Bootstrap Results

4.3.1 Sampling Distribution of the Treatment Effect Estimate Based on the Bootstrap

The nonparametric bootstrap analysis based on the final model is more appealing in this study as the normality assumption about the distributions of random effects is not satisfied. This approach relieves one from questionable assumptions, and also mathematical derivations are avoided at the cost of computing. By using an iterative procedure, the bootstrap is expected to yield a bias-corrected estimate for the treatment effect (Hutchinson, Morrison, & Felgate, 2003).

The bootstrap sampling distribution of the treatment effect estimate is displayed in Figure 4.18. It can be seen that the shape of the distribution is very close to normal, which can be accounted for by the Central Limit Theorem. In fact, the normal Q-Q plot in Figure 4.19 also suggests that this distribution is normal since the absolute majority of the points fall on the reference line with a slight departure for only a few points at the two ends.

Furthermore, based on the Law of Large Numbers, the bootstrap estimated mean theoretically converges in probability to the population mean as the number of bootstrapped samples approaches infinity. In this bootstrap study there are 1000 independent and identically distributed samples, so the mean of the sampling distribution in Figure 4.18 is expected to be very close to the actual population average.
Figure 4.18: The Sampling Distribution of the Estimate of the READ 180 Treatment Effect from the Bootstrap

Figure 4.19: Normality Q-Q Plot of the 1000 Estimates of Treatment Effect based on Independent Bootstrapped Samples
Table 4.19 presents the mean, standard error, and confidence intervals produced by the SAS PROC MIXED procedure based on the normality assumption and those provided by the bootstrap analysis. There is a very small difference between the two values for the estimated mean treatment effect (6.2364 vs. 6.0574), with the bootstrap estimated one slightly larger. The bootstrap analysis also provides the median of the sampling distribution of the treatment effect estimate, and it is quite close to the mean of that distribution. Moreover, the standard error estimated by the bootstrap is a little bigger than that outputted by SAS PROC MIXED. It can also be seen that the confidence intervals provided by these two procedures do not differ much, with the bootstrap estimated confidence intervals larger in width. Also, the confidence intervals based on both procedures support that there is a significant program impact of the READ 180 on the reading level of low-performing incarcerated youth, since they do not cover the null point.

In sum, the results yielded by the bootstrap analysis are very similar to those given by SAS. The SAS PROC MIXED estimates are not considerably biased.
according to the comparison. Therefore, they largely support the robustness of the mixed-effects modeling to the violation of the normality in this study.

4.3.2 Power via the Bootstrap and the Analytical Approximation in SAS

Power calculation based on the bootstrap simulation is described in Section 3.5.2, while the SAS PROC MIXED procedure employs a simple analytical approximation to the noncentrality parameter of the $F$-distribution to compute approximate power for mixed models for an existing data set.

Power curves produced from the above two procedures are shown in Figure 4.20, with the significance level plotted along the horizontal axis and power along the vertical axis, keeping other factors unchanged. It can be seen that the bootstrap estimated powers and SAS estimated ones at given Type I errors are quite close, with the former consistently a little higher than the latter.

According to Table 4.20, the power of testing the treatment effect at the default $\alpha$ level is pretty descent since it is around .85 based on both approaches. When $\alpha$ is set at .01, power is not so satisfactory since it is below .70.

<table>
<thead>
<tr>
<th>Power</th>
<th>Type I Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.01</td>
</tr>
<tr>
<td>SAS PROC MIXED</td>
<td>.655</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>.676</td>
</tr>
</tbody>
</table>

Table 4.20: Power Comparison between the Bootstrap and the SAS PROC MIXED
Figure 4.20: Power Curves Produced by the Bootstrap and the SAS PROC MIXED

4.4 Monte Carlo Simulation Results

Since a series of Monte Carlo studies are conducted, the results will be presented in the following order. The simulation with unbalanced data, sample size of 300, treatment effect magnitude set to the bootstrap estimated mean, and 5 points in time, will be presented first as its results are expected to be quite comparable with the bootstrap analysis with a sample size of 272. Then other simulations with different sample sizes, magnitudes of treatment effect, numbers of time points, and data patterns will be organized for power comparison purposes.
4.4.1 Comparison with the Bootstrap Analysis

4.4.1.1 Statistical Properties of the Probability Distribution of the Intervention Effect Estimate

The probability distributions of the treatment effect estimate based on the bootstrap and the Monte Carlo with comparable design factors are displayed in Figure 4.21. It can be seen that the shape of both distributions is approximately normal, with very similar central tendency and spread.

Table 4.21 lists the means, medians, standard errors, and confidence intervals provided by the bootstrap and the Monte Carlo simulation. The difference in the estimated mean treatment effect between the two approaches is quite small (6.1440 vs. 6.2364). Recall that in Table 4.19 the SAS PROC MIXED gives an estimate of 6.0574. Thus the mean provided by Monte Carlo is somewhere in between the other two values. Again, there is a very small difference between the two estimated medians (6.1661 vs. 6.2501). With both approaches, the median is very close to the mean estimate of the distribution.

In addition, the standard error estimated by Monte Carlo is quite similar to that from the bootstrap analysis, only a little smaller (2.1355 vs. 2.1506). Both standard errors are slightly larger than that given by SAS (2.0163). As for the confidence intervals produced by the two approaches, they do not differ substantially, with the 99% CI given by the Monte Carlo a little narrower than the corresponding one based on the bootstrap. Moreover, the information provided by these confidence
intervals confirm that the READ 180 program has a significant treatment effect on the reading proficiency of low-achieving incarcerated youth, since even the two 99% CIs do not include zero.

The above facts indicate that the Monte Carlo simulation design is reasonably reliable, and the slight variation in statistical properties of the probability distributions of the treatment effect estimate might be attributed to the small difference in sample sizes (300 vs. 272) or random error. Overall, one is more confident that the READ 180 program should be effective for the target population.

Figure 4.21: Probability Distributions of the Estimate of the READ 180 Treatment Effect via the Bootstrap and the Monte Carlo with Comparable Design Factors
<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Std. Err.</th>
<th>95% CI</th>
<th>99% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bootstrap</td>
<td>6.2364</td>
<td>6.2501</td>
<td>2.1506</td>
<td>[2.1279, 10.4793]</td>
<td>[0.8495, 11.9137]</td>
</tr>
<tr>
<td>MCS</td>
<td>6.1440</td>
<td>6.1661</td>
<td>2.1355</td>
<td>[1.9720, 10.3724]</td>
<td>[1.0135, 11.3989]</td>
</tr>
</tbody>
</table>

Table 4.21: Mean, Median, Standard Error, and Confidence Intervals Produced by the Bootstrap and the Monte Carlo with Comparable Design Factors

4.4.1.2 Power Comparison

The power analysis by Monte Carlo is unique in that both the null and alternative effect distributions can be simulated for precise calculation of the probability of rejecting the null hypothesis when the alternative is true.

As described previously, the distribution under the null hypothesis is simulated by setting the treatment effect estimate to zero while keeping other factors unchanged. Based on this null distribution, one can calculate the \( p \)-value for an alternative magnitude of treatment effect and also plot the relevant power curve.

Figure 4.22 depicts the simulated null distribution with a sample size of 300 and 5 points in time. It also displays the alternative distribution with the treatment effect set to the bootstrap estimated mean (i.e., 6.2364) and other equal design factors. It is observed that the two distributions are similar in shape as expected. The only apparent difference is the shift in the location of mean estimate.
The power curve for the alternative treatment effect of 6.2364 is therefore plotted in Figure 4.23. The power curve produced from the bootstrap analysis is also presented in the same figure for comparison. It can be seen that the two power curves are very close to each other, indicating minor difference.

According to Table 4.22, the power of testing the treatment effect at the significance level of .05 is high enough, both approaches yielding values above .86. When the significance level is lowered to .01, power given by Monte Carlo simulation reaches .70 while that from the bootstrap analysis is around .68.

Overall, power based on the Monte Carlo is a little higher. This slight increase in power is not surprising, given the fact that the sample size of 300 used in the corresponding simulation is a little larger than that used for resampling in the bootstrap (272), i.e., the size of the original data set.

<table>
<thead>
<tr>
<th>Power</th>
<th>Type I Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.01</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>0.676</td>
</tr>
<tr>
<td>Monte Carlo</td>
<td>0.703</td>
</tr>
</tbody>
</table>

Table 4.22: Power Comparison between the Bootstrap and the Monte Carlo with Comparable Design Factors
Figure 4.22: Monte Carlo Simulated Null Treatment Effect Distribution vs. Alternative Treatment Effect Distribution for Sample Size of 300 and 5 Time Points

Figure 4.23: Power Curves Produced by the Bootstrap and the Monte Carlo with Comparable Design Factors
4.4.2 Influence of Magnitude of Treatment Effect on Power with Unbalanced Data

After comparing the results from the bootstrap and the Monte Carlo simulation with similar conditions, the focus now is on power comparisons by varying different design factors, usually one at a time.

The magnitude of the treatment effect is a most important factor that is considered in impact studies by researchers. As mentioned before, with the underlying population being low-achieving incarcerated youth, a more sensible way to changing the levels of the treatment effect is based on the estimates from the bootstrap sampling distribution instead of the commonly used Cohen’s effect size, because the former represents more plausible values for this factor.

Here the Monte Carlo simulations are carried out by varying the magnitudes of the treatment effect while holding other factors constant. In other words, the sample size is still set to 300, and the number of time points is 5 still, with unbalanced data pattern. Under these conditions, the null effect distribution has been simulated already. Thus only alternative distributions at different magnitudes of treatment effect need to be simulated.

According to the design specified in Section 3.6.5.1, the bootstrap estimated .025, .25, .75, and .975 quantiles of the treatment effect distribution will be used here as four of the five magnitude levels. Due to the fact that the estimated mean and the median (the .50 quantile) in the bootstrap analysis is very close, the bootstrap estimated mean is used here as another level for effect magnitude.
Figure 4.24: Simulated Distributions of the Treatment Effect Estimate at the Null, .025 Quantile, .25 Quantile, Mean, .75 Quantile, and .975 Quantile based on the Bootstrap

Figure 4.25: Power Curves for the Treatment Effect at the .025 Quantile, .25 Quantile, Mean, .75 Quantile, and .975 Quantile at Different Type I Error Rates
Figure 4.26: Power Pattern at the .025 Quantile, .25 Quantile, Mean, .75 Quantile, and .975 Quantile of the Treatment Effect for $\alpha$ at .01 and .05

Figure 4.27: Power Pattern at Different Effect Sizes for $\alpha$ at .01 and .05
Note: * The bootstrap estimated mean is very close to the median, namely the .50 quantile.

Table 4.23: The Corresponding Cohen’s Effect Sizes for the Bootstrap Estimated Quantiles of the Treatment Effect Distribution

<table>
<thead>
<tr>
<th>Quantile</th>
<th>.025</th>
<th>.25</th>
<th>Mean*</th>
<th>.75</th>
<th>.975</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate Magnitude</td>
<td>2.1279</td>
<td>4.7851</td>
<td>6.2364</td>
<td>7.6537</td>
<td>10.4793</td>
</tr>
<tr>
<td>Cohen’s Effect Size</td>
<td>0.006</td>
<td>0.031</td>
<td>0.053</td>
<td>0.080</td>
<td>0.149</td>
</tr>
</tbody>
</table>

Figure 4.24 displays the null effect distribution and the other five alternative distributions with treatment effect set at different magnitudes based on the bootstrap estimates. It can be seen that all these distributions are similar in shape. This is because they are simulated based on same design factors other than the effect magnitude.

Power of the test of the treatment effect at different magnitudes can then be examined using these simulated distributions. This analysis is particularly helpful since the real treatment effect could be lower or higher than the bootstrap estimated mean. Corresponding power curves are plotted in Figure 4.25.

Specifically, power is extremely low when the effect magnitude is equal to the bootstrap estimated .025 quantile (i.e., 2.1279). This is not surprising since this magnitude is near to the null effect. Therefore, it is very unlikely that a trivial treatment effect such as the .025 quantile can be detected with a sample size of 300. For the effect magnitude at the .25 quantile (i.e., 4.7851), power improves...
considerably but is still not high enough as the value is below .80 at the $\alpha$ level of .05. However, when the treatment effect is set to the bootstrap estimated mean, namely 6.2364, power of the test is sufficiently high as discussed before. When the effect magnitude is increased to the .75 quantile (i.e., 7.6537), power is very high even if the significance level is set at a more stringent level such as .01. Also as shown in Figure 4.25, if the effect magnitude is set at the .975 quantile (i.e., 10.4793), power becomes extremely high even when the significance level approaches to zero.

The general pattern of these power curves shows that as expected, power increases as the magnitude of the treatment effect gets larger. This pattern can also be detected in Figure 4.26 in which the trend of power improvement is displayed respectively at Type I error rate of .05 and .01. Table 4.23 presents the values calculated for Cohen’s effect size at the five corresponding effect magnitudes. It can be seen that the .025 quantile relates to a very small effect size, while the .0975 quantile refers to a medium effect size based on the Cohen’s rule. The magnitudes in between (i.e., the .25, .75 quantiles, and the mean) all indicate small effect size. Again as seen from Figure 4.27, power at given Type I error rate will increase as the effect size increases.

In conclusion, given a sample size around 300 and with other factors kept constant at the same time, it would be difficult to detect the treatment effect if the magnitude falls below the .25 quantile. Therefore, it may help researchers by investigating the influence of sample size on power, especially when the treatment effect is fairly low.
4.4.3 Influence of Sample Size on Power with Unbalanced Data

4.4.3.1 When the Magnitude of Treatment Effect is Equal to the Bootstrap Estimated Mean

Based on the above discussion, a study that naturally follows is to explore whether sample size has an effect on power when testing the treatment effect. The alternative magnitude of the effect will first be set at the bootstrap estimated mean (i.e., 6.2364), since it is the most plausible value based on previous analyses. Again, unbalanced data with 5 time points are simulated here for this investigation. Only the sample size will be respectively set at 150 and 600, as compared to the previous size of 300.

To calculate power at the three levels of sample size, three sets of corresponding null and alternative effect distributions based on different sample sizes are needed. These three sets of simulated distributions are shown in Figure 4.28. It is noted that as the sample size gets larger, the shape of the distribution becomes more leptokurtic. In theory, standard error of the sampling distribution will decrease as the sample size increases.

Power curves corresponding to three sample sizes are depicted in Figure 4.29. When the sample size is 150, power is not satisfactory at the default Type I error rate. The power curve for sample size of 300 has been shown before, and it is known that power of detecting the treatment effect at the magnitude of 6.2364 is high enough. As the sample size increases to 600, power is quite high even at a more stringent $\alpha$
level such as .01. Therefore, power will be improved as sample size increases. This general trend can also be observed in Figure 4.30 in which the pattern of power increment is shown at the significance levels of .05 and .01 respectively. It can be concluded that at a sample size as small as 150, one might not be able to detect a treatment effect at the magnitude of 6.2364 with unbalanced data at 5 time points.

In addition, as indicated in the previous section, when the actual treatment effect is smaller than 6.2364, even a sample size of 300 may not be large enough to detect it. As a result, another two simulations are carried out with a sample size of 600 and by setting the magnitude of treatment effect to the bootstrap estimated .025 and .25 quantiles respectively while holding other design factors unchanged.

Figure 4.28: Simulated Distributions of the Treatment Effect Estimate at the Null and the Bootstrap Estimated Mean with Unbalanced Data at 5 Points in Time for Sample Size of 150, 300, and 600
Figure 4.29: Power Curves for Sample Size of 150, 300, and 600 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated Mean

Figure 4.30: Power Pattern at Sample Size of 150, 300, and 600 for \( \alpha \) at .01 and .05
4.4.3.2 When the Magnitude of Treatment Effect is Equal to the Bootstrap Estimated 0.025 Quantile

If the magnitude of the treatment effect is set to 2.1279, namely the .025 quantile from the bootstrap analysis, power curves for detecting the effect of this large are shown in Figure 4.31 for sample size of 300 and 600 respectively. Remember that the simulation is based on unbalanced data and 5 points in time. It can be seen that power is still very low even when the sample size is raised to 600. Due to unaffordable computational expense, the sample size of 1200 is not conducted for power comparisons, but it is speculated that this size may not benefit power to a satisfactory degree at the default Type I error, given such a small treatment effect.

Figure 4.31: Power Curves for Sample Size of 300 and 600 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated .025 Quantile
4.4.3.3 When the Magnitude of Treatment Effect is Equal to the Bootstrap Estimated 0.25 Quantile

Now the magnitude of treatment effect is placed at 4.7851, namely the bootstrap estimated .25 quantile. Again, the simulation is based on unbalanced data and 5 points in time. Sample size is set to 300 and 600 respectively for comparison. Power curves corresponding to these two sample sizes are displayed in Figure 4.32. It can be observed that while a sample size of 300 may not be sufficient to detect the treatment effect as large as 4.7851 at the $\alpha$ level of .05, a sample size of 600 increases power to around .85 given the same Type I error rate. Even when $\alpha$ is set to .01, power for a sample size of 600 is still as high as .80.

![Power Curves for Sample Size of 300 and 600 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated .25 Quantile](image)

Figure 4.32: Power Curves for Sample Size of 300 and 600 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated .25 Quantile
4.4.4 Influence of Number of Time Points on Power with Unbalanced Data

Another study of interest is to investigate if varying the number of time points will affect power. The study will be first carried out with unbalanced data, to be consistent with all the previous Monte Carlo simulations. The magnitude of treatment effect is again set to the bootstrap estimated mean (i.e., 6.2364), and the sample size will be placed at 300. The number of time points will be set to 6 and 7 respectively, to be compared with the original 5.

Again, three sets of corresponding null and alternative effect distributions based on different points in time are simulated for the power calculations. These simulated distributions are shown in Figure 4.33. It can be seen that the alternative distributions with different time points are very close to each other, and the same pattern is observed for the three null distributions. Thus power for the three different numbers of time points is not expected to differ much. This is confirmed by the power curves plotted in Figure 4.34. It is noted that power is not appreciably improved by increasing points in time. Figure 4.35 also shows the same pattern in that the two power curves at different number of points in time are nearly flat at both \( \alpha \) levels of .05 and .01.

This fact is largely due to the high drop-out rate observed previously. As shown in Figure 4.35, it is predicted that only a small number of incarcerated youth will remain in the program after one year (i.e., 5 points in time including the baseline). Therefore, with realistic unbalanced data, the influence of number of time points (beyond 5) on power would be minor.
Figure 4.33: Distributions of the Treatment Effect Estimate at the Null and the Bootstrap Estimated Mean for Time Points of 5, 6, and 7 with Unbalanced Data

Figure 4.34: Power Curves for Time Points of 5, 6, and 7 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated Mean with Unbalanced Data
4.4.5 Power Comparisons for Balanced and Unbalanced Data

It is noticed that increasing the number of time points does not improve power substantially in highly unbalanced data with remarkable attrition rate, since essentially there is little increase in the number of level-1 units. Typically with balanced data, the influence of number of points in time on power is expected to be more obvious. For theoretical purposes, balanced data will also be simulated to be compared with unbalanced data and to explore if increasing points in time will affect power.

First, balanced data with 5 time points, sample size of 300, and treatment effect equal to 6.2364 will be compared to unbalanced data with the same design.
factors. The simulated distributions are shown in Figure 4.36. As expected, the null and alternative distributions for balanced data are more leptokurtic than those for unbalanced data, largely because the number of level-1 units for the former is much larger than that for the latter.

It is known that the balanced design has an advantage over the unbalanced one with other conditions equal. Figure 4.37 also shows that power for balanced data is consistently higher than that for unbalanced data at given Type I error rates. It can be seen that when $\alpha$ is set to .01, power for unbalanced data is merely .70, whereas that for balanced data is over .90. Although in reality balanced data are highly unlikely to be obtained under the present study context, the degree of unbalance caused by many factors should be minimized whenever possible.

Figure 4.36: Distributions of the Treatment Effect Estimate at the Null and the Bootstrap Estimated Mean with Balanced and Unbalanced Data
Figure 4.37: Power Curves for Balanced and Unbalanced Data at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated Mean

4.4.6 Influence of Number of Time Points on Power with Balanced Data

At this time one may proceed to further investigate the influence of number of points in time on power with balanced data. Again, with the treatment effect magnitude equal to 6.2364, the number of time points will be set to 6 and 7 in addition to 5.

The sample size will be placed at 300 first to compare relevant results with those for unbalanced data. The simulated distributions and power curves are displayed in Figure 4.38 and Figure 4.39 respectively. It can be seen that power is
extremely high at the default $\alpha$ level, and the consequence of increasing points in time is not noticeable given the magnitude of treatment effect and sample size. As the Type I error rate approaches zero, however, the influence of points in time can be observed. When the number of time points is increased by 1 in the balanced design, power is appreciably higher. If a stringent $\alpha$ level (e.g., .01) is adopted, a similar power pattern can be seen in Figure 4.40 also.

With a balanced design, it is expected that power can still be sufficient with a more economical sample size. Thus a sample size of 150 is used next for another simulation for balanced data while other design factors are kept unchanged. The simulated distributions and power curves are respectively shown in Figure 4.41 and Figure 4.42. It is noticed that at the significance level of .01, power for the balanced data with the sample size of 150 is around .70, but it is improved to above .80 when the number of time points increases by 1, and even higher with one more time point. This power pattern is also observed in Figure 4.43.

In general, the above simulations show the advantage in design for balanced data. It is speculated that the benefit of increasing the number of time points on power in a longitudinal study with a balanced design may be more apparent with an even smaller sample size, and possibly with a smaller magnitude of treatment effect.
Figure 4.38: Distributions of the Treatment Effect Estimate at the Null and the Bootstrap Estimated Mean for Time Points of 5, 6, and 7 with Balanced Data and Sample Size of 300

Figure 4.39: Power Curves for Time Points of 5, 6, and 7 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated Mean with Balanced Data and Sample Size of 300
Figure 4.40: Power Pattern at Time Points of 5, 6, and 7 for \( \alpha \) at .01 and .05 with Balanced Data and Sample Size of 300

Figure 4.41: Distributions of the Treatment Effect Estimate at the Null and the Bootstrap Estimated Mean for Time Points of 5, 6, and 7 with Balanced Data and Sample Size of 150
Figure 4.42: Power Curves for Time Points of 5, 6, and 7 at Different Type I Error Rates with the Treatment Effect at the Bootstrap Estimated Mean with Balanced Data and Sample Size of 150

Figure 4.43: Power Pattern at Time Points of 5, 6, and 7 for $\alpha$ at .01 and .05 with Balanced Data and Sample Size of 150
CHAPTER 5

SUMMARY, DISCUSSION, AND CONCLUSION

This chapter starts by summarizing the findings of this study. Following the summary, implications of the major findings are discussed. Limitations of the study and recommendations for future research are presented at the end.

5.1 Summary and Discussions

This study was conducted with two major focuses. One was to investigate if the READ 180 program had a significant influence on the reading proficiency of low-performing incarcerated youth, and the other was to provide statistical properties for the sampling distribution of the estimate of the READ 180 treatment effect on low-performing incarcerated youth as well as methodologically guiding future longitudinal research with power analysis.

The study was carried out based on a longitudinal experimental design in which the eligible incarcerated youth were randomly assigned to either the READ 180 program or a comparison group being instructed by a traditional reading program. The course of investigation lasted for a school year during which the subjects were measured for their reading proficiency by the Scholastic Reading Inventory (SRI) prior to the treatment and again at the end of each of the four terms.
Demographic information was also collected together with another two baseline measures, namely the ReadCAT and MathCAT. The target population of the study included all low-performing incarcerated youth in a large mid-western state. The accessible population in this study included all the low-achieving incarcerated youth in the state who were determined to be eligible for the READ 180 program from the autumn of 2006 to the summer of 2007. The study sample was selected based on the criterion that they received at least two terms of instruction. No significant difference was detected for the demographics and the baselines measures of the study participants and non-participants.

Mixed-effects modeling was adopted to analyze the sample data due to its methodological advantages. Competing hypotheses about the mean response over time and the covariance structure for the residuals were proposed for model fitting. Using a two-step approach, the final model was selected based on fit indices, the interpretability, and the principle of parsimony. This empirically-verified model was then employed in the simulation studies to fulfill the second research purpose. The bootstrap technique was applied to the original sample first, followed by the Monte Carlo simulations based on the information from the accessible population.

A summary of the findings regarding each of the six research questions is presented in the following. Since the dominating majority of the accessible population was male, generalizing these findings to female incarcerated youth should be cautioned.
1. Does the READ 180 program have a significant impact on the reading level of low-achieving incarcerated youth over time, when comparing the experimental group with the comparison group being instructed with the traditional educational reading program on reading achievement and growth?

Based on the study sample, it was found that the READ 180 program exerted a significant positive impact on the reading proficiency of low-achieving incarcerated youth as compared to the comparison group, although it took time for the subjects to display remarkable growth. In fact, the targeted experimental group lagged behind during the former half of the school year although its mean performance on the SRI was steadily improved over time. This was in accordance with the Scholastic’s (2004) claim about enough treatment exposure. The subjects were not expected to adjust to the newly implemented program in a short time.

Surprisingly, the average score of the comparison group increased in the beginning, especially after the first 10-week instructional block, but decreased substantially after the third term. In regard to juvenile delinquents, the longer they are incarcerated, the more problems they typically have. Thus for the comparison group who did not receive a quality reading program, it was expected that there would be tapering off in their reading scores after a certain time period.

The mixed-effects modeling of the sample data resulted in a significant positive estimate of 6.06 for the READ 180 treatment effect over time with a quadratic trend. In other words, the subjects exposed to READ 180 demonstrated a
significant accelerated reading growth over time, when compared to their counterparts. The reading achievement gap between these two groups widened greatly over time. On average, a typical READ 180 program participant outperformed a student in the comparison group by about 100 lexile points after one school year. According to Meta-Metrics, the expected growth for a typical ninth grader following a year of instruction by the READ 180 is around 45 to 55 lexile points. Given this fact, it seemed that the reading growth demonstrated by the low-performing incarcerated youth in the experimental group in this study was exceptional. Therefore, the READ 180 program worked quite well for the low-performing incarcerated youth in the state.

Nevertheless, one needs to bear two points in mind. The first was that tremendous inter-individual and intra-individual variability was observed for the reading level in both groups. The second was that after four terms of treatment, a typical student in the READ 180 group was still not reading at grade level, although significant reading progress had been achieved.

Moreover, it is critical that the randomization process in randomized controlled trials is valid so that causal conclusions can be drawn. In this case, the baseline equivalence between the experimental group and the comparison group was additionally verified by including the treatment predictor together with other covariates when modeling the initial reading status. The results indicated that as expected, the experimental group and the comparison group did not differ
significantly in terms of the initial reading status. This provided one with more confidence in the validity of the randomization process in the study.

2. Can the variability in the initial reading status of low-performing incarcerated youth be attributed to subject characteristics including age, gender, race/ethnicity, disability status, grade level, special education status, school mobility, and other baseline assessments including Reading and Math of the California Achievement Test (CAT)?

The results suggested that the initial reading status was only influenced by a subject’s age, disability status, and the baseline ReadCAT scores. Specifically, on average a subject’s initial reading score was expected to increase 33.46 lexile points with one year’s increase in age, and 38.27 points with one unit increase in the baseline ReadCAT score. The average difference in initial reading between low-achieving incarcerated youth with disability and those without disability was 51.87. Generally these findings were in accordance with one’s anticipation regarding the influences of the three covariates. The more mature the reader, the higher reading level he possessed. The ReadCAT score also appeared to be a valid measure of reading proficiency as expected. With regard to disability status, it was known that around half of the incarcerated youth in this study had various disabilities. The Coalition for Juvenile Justice (2001) indicates that at the national level there are between 70 and 87 percent of incarcerated youth suffering from learning or emotional
disabilities. These disabilities considerably interfere with their education and affect their proficiency in reading.

3. Can the variability in the reading growth rate of low-performing incarcerated youth be accounted for by subject characteristics including age, gender, race/ethnicity, disability status, grade level, special education status, school mobility, and other baseline assessments including Reading and Math of the CAT?

Based on the findings, the MathCAT turned out to have a significantly positive effect on the rate of reading growth with a linear trend. To be specific, one unit increase in the baseline MathCAT score was expected to improve an average of 5.22 lexile points constantly over each 10-week instructional period. Research suggests that reading achievement can predict math outcomes (Fletcher, 2005). In this study, it was also known that the baseline MathCAT and ReadCAT had a significant correlation of 0.65. It seemed that MathCAT could explain an appreciable amount of variance in the constant rate of change during the course of investigation. Those with higher math ability improved their reading level very stably over time.

In addition, the grade level had a significantly negative impact on the reading growth rate with a quadratic trend. In other words, at higher grade levels, the reading growth was expected to decelerate over time. On average, with one level’s advance in current grade, a low-performing incarcerated youth was expected to slow down the reading growth by 12.48 lexile points after one school year. According to Meta-Metrics, the expected growth for typical students receiving the READ 180 instruction
slows precipitously as they advance in grade levels. Here this finding suggested that a similar decelerating trend had been detected for the low-achieving incarcerated youth in the state.

It was noted that no significant results were found for the following three covariates: institutions, school mobility, and race/ethnicity. This indicated that none of these covariates could account for the variability in the initial reading status or the rate of reading growth of the targeted low-achieving incarcerated youth. Normally race/ethnicity is an important predictor of academic achievement and literacy skills (Daggett & Hasselbring, 2007). Here it seemed that the impact of this risk factor was mitigated among low-performing incarcerated youth. The non-significant findings regarding institutions and school mobility might imply that the implementation fidelity of the READ 180 Instructional Model was similar across different institutions. Note also that institutions were confounded with gender, and the impact of the latter could not be examined with only a few female youth available.

4. What are the statistical properties of the probability distribution of the READ 180 intervention effect estimate, in terms of mean, standard error, and confidence intervals?

Based on the bootstrap analysis, the estimated mean for the READ 180 intervention effect was 6.24, with a standard error of 2.15. The 95% bootstrap confidence interval was [2.1279, 10.4793] and the 99% one was [0.8495, 11.9137]. These results were only slightly different from the estimates provided by the mixed-effects modeling of the empirical data set in SAS. Note that although the bootstrap
analysis characterized the sampling distribution of the intervention effect estimate based on the original sample of data only, it was capable of providing bias-corrected estimators using an iterative procedure.

Based on the Monte Carlo simulation with comparable design factors, the estimated mean for the READ 180 treatment effect was 6.14, with a standard error of 2.14. The 95% confidence interval provided by the Monte Carlo was [1.9720, 10.3724] and the 99% one was [1.0135, 11.3989]. It seemed that the Monte Carlo design was very reliable, given the fact that the results provided by the bootstrap and the Monte Carlo were quite similar.

Overall, the results based on the simulation studies were consistent with those provided by the pilot study based on the original sample data, supporting that the READ 180 is an effective intervention that can help accelerate reading growth for low-achieving incarcerated youth over time.

In addition, recall that the analysis of the study sample showed that both the univariate normality assumption for the level-1 residuals and the multivariate normality assumption for the level-2 random effects were violated. The results based on the Monte Carlo simulation largely supported the expected robustness of the mixed-effects modeling to the violation of the normality assumption (Zhang, 2005), since the results based on simulated data with normally distributed residuals and random effects were quite consistent with those given by the actual sample. Hence the results yielded by the mixed-effects modeling of the original data set were trustworthy.
Furthermore, data simulation via the Monte Carlo approach was based on the information of all the 724 low-achieving incarcerated youth. Thus the consistent findings from the Monte Carlo simulation with comparable design factors implied that the study sample was representative of the underlying population, even though the subjects were not selected by use of random sampling. The generalization of findings to the target population therefore could be considered as appropriate.

5. What is power of the statistical test of the READ 180 intervention effect, based on both the available sample from a pilot study perspective and the simulations?

The probability of detecting a real READ 180 intervention effect was quite decent in this case. With a Type I error rate of .05, the analytical approximation approach provided by SAS PRCO MIXED provided an estimate of .848 for the power, while the bootstrap and the Monte Carlo approaches yielded 0.864 and 0.879 respectively. The consistent results suggested that the statistical power in this study was sufficient to detect a real READ 180 treatment effect at the default significance level. If the $\alpha$ was lowered to a more stringent level of .01, however, the power was not satisfactory enough.

2. Based on Monte Carlo studies, how does this power change in the context of longitudinal research when varying magnitude of treatment effect, sample size, number of time points, Type I error rate, and data pattern?
Under different pre-specified scenarios, a series of Monte Carlo studies were carried out to examine the issue of power in similar longitudinal research. The realistic unbalanced data pattern was focused on for power comparisons. When simulating data for the covariates, it was assumed that the correlations among the covariates based on the information of all the 724 eligible youth hold in the underlying population. It is wise to verify this assumption by examining the same correlations as more data are gathered from low-performing incarcerated youth in the state in the future.

The results demonstrate that the magnitude of treatment effect has an apparent influence on power. As expected, power increases as the effect size gets larger. While keeping other design factors unchanged, power is extremely high even at very stringent Type I error rates when the effect magnitude is as large as the bootstrap estimated .975 quantile. In contrast, power is extremely low when the effect magnitude is equal to the bootstrap estimated .025 quantile which is close to the null. For the effect magnitude at the bootstrap estimated .25 quantile, power improves considerably up to around .70.

The influence of sample size on power is also quite obvious. When the magnitude of treatment effect is set to the bootstrap estimated mean, power is satisfactory with a sample size of 300 but not 150. In addition, increasing the sample size to 600 for a treatment effect as low as the bootstrap estimated .025 quantile does not enhance power greatly, but for a magnitude around the bootstrap estimated .25 quantile, doubling the sample size brings power up to a satisfactory level.
With regard to the number of time points, it seems that this design factor does not substantially influence power under the unbalanced design when the attrition rate is fairly high. For theoretical balanced data with a reasonably small sample size, however, increasing the number of points in time may show a noticeable effect. In addition, a balanced data pattern displays higher power when compared to unbalanced data, keeping other design factors constant.

5.2 Implications

This study empirically examined the program impact of the Scholastic READ 180 for low-performing incarcerated youth in a large mid-western state. Based on consistent results, it is concluded that the READ 180 program is effective for low-achieving incarcerated youth in this particular state. It is also believed that the target population can accelerate their reading growth over time following enough exposure to the READ 180 instruction, which may eventually lead to the closing of the reading achievement gap.

In regard to the generalizability of the findings, the nonrandom sample used for the analysis did not cause much concern, not only because comparison of various characteristics of the participating and non-participating group did not indicate any significant differences, but also due to the fact that this limitation was ameliorated with the Monte Carlo approach. Nonetheless, the findings should be interpreted with caution in that they may only generalize to male low-achieving incarcerated youth in the state. It was noted that the study sample were mostly males, so one should be
cautious when making conclusions regarding female incarcerated youth. With limited information, one is not positive how validly the results can be applied to the female group. For a population of low achievers with similar characteristics as the low-performing incarcerated youth in the state, however, the results obtained in this study about the READ 180 program may apply as well. The READ 180 appears to be a promising program to help to address the urgent national crisis in adolescent reading achievement, particularly for incarcerated youth.

For low-performing incarcerated youth in particular, a quality reading program that can engage them in effective reading practices is very crucial. As discussed before, those juvenile delinquents who are detained for a longer time tend to have more problems. With ineffective programs they normally achieve little progress or even regress upon certain point, just as the pattern displayed by the youth in the comparison group in this study. In contrast, the READ 180 seemed to be able to mitigate certain risk factors in that the experimental group demonstrated accelerated reading growth over time. It seemed that the READ 180 was qualitatively different from the traditional English class, although the subjects in the comparison group made appreciable progress in reading compared to their baseline SRI scores.

In addition, the length of time it takes for the READ 180 recipients to demonstrate remarkable growth should be taken into account when making policy-related decisions. It seems that although the READ 180 is a high quality program, the time that it takes for low-performing incarcerated youth to adjust to this newly implemented program is no less than half of the school year. Thus sufficient
treatment amount should be allowed for the youth exposed to the READ 180 to demonstrate remarkable reading growth.

Moreover, the typical youth in the study was still not reading at grade level while significant reading progress had been achieved. The READ 180 program is expected to help these youth achieve literacy skills required in school and in life. As the treatment program continues to mature, it would be interesting to see if these struggling readers eventually achieve grade level status in their reading.

Also given that the expected growth decelerates substantially as students advance in grade levels, the READ 180 instruction should start from lower grade levels whenever and wherever possible to achieve better intervention outcomes.

It was also noted that even though the READ 180 demonstrated significant positive impact on the reading level of low-achieving incarcerated youth, the effect size measures were still small based on Cohen’s criteria. Only the bootstrap estimated .975 quantile approached a medium effect size. When further converting effect size measures into the WWC (2007b) Improvement Index, the bootstrap estimated mean effect only corresponded to 2.11 percentile points, which indicated that the average student in the comparison group would have gained at least 2.11 percentile points in achievement if exposed to the READ 180 intervention. The average effect might not appear to be so impressive. This was largely due to the unusually large inter-individual and intra-individual variability shown in the outcome SRI scores. Typical standard error on the SRI is around 55 lexile points according to Meta-Metrics, whereas over 200 points were detected in this study consistently across
different measurement occasions. One may be especially concerned about huge variance in impact studies since significant effects are less likely to be detected.

In fact, the methodological considerations in program evaluation of the READ 180 for the target population are central. Preliminary analyses using classical approaches such as repeated measures ANOVA did not detect any significant treatment effect for the experimental group. Mixed-effects modeling was successful in detecting this treatment effect with statistical control of multiple covariates in the same analysis.

With regard to statistical power, if the actual magnitude of the intervention effect is smaller than the estimated mean via the bootstrap and the Monte Carlo approaches, increasing sample size may help detect this effect as long as it is not too trivial. On the other hand, if the actual treatment effect is larger than the estimated average, e.g., the bootstrap estimated .75 quantile or even the .975 quantile, sufficient power can be achieved even with a more economical sample size. It also seems that when the attrition rate is fairly high, increasing the duration of the investigation may not appreciably enhance power.

5.3 Limitations

The following limitations of this study should be noted:

1. The influence of gender on the variability in the initial reading status and the reading growth rate was not examined effectively due to an extremely small number of female incarcerated youth in the available
data. The generalizability of the study findings was also limited when considering the female group.

2. Other variables that were not included in the mixed-effects modeling might be better predictors for the initial reading proficiency and the growth rate of low-performing incarcerated youth. For instance, the type of felony the youth committed, and their family social-economic status, the medication level, etc. are possible factors that could explain the variability. However, these pieces of information are not available at this time.

3. Given the size of the study sample, more complex mixed-effects models could not be considered due to estimation problems.

4. Regarding the identification of outlying cases, only the rule of thumb was employed for analysis since the number of items a student answered during a test and the length of time that a student took to complete the test were both unknown to the researcher.

5. Without any item level data for the SRI, the validity and reliability of the computer-adaptive test could not be examined. There was a possibility that this outcome measure was not as accurate as it was thought to be. Relying on a single measure for reading proficiency was not desirable.
5.4 Recommendations for Future Research:

In light of the major findings and limitations of this study, the following recommendations may be useful in conducting similar studies in the future:

1. Efforts should be devoted to investigating the possible gender effect on the reading achievement and growth of low-performing incarcerated youth when the relevant information is gathered for an adequate number of female incarcerated youth.

2. Meta-Metrics claims that a negative growth slope in the SRI score is highly unlikely for the READ 180 participants. However, there were approximately a third of subjects displaying negative growth slopes. Future research may focus on the prediction of negative reading growth slopes with post CAT scores. In addition, one can study the characteristics of those with negative slopes in an effort to detect possible patterns or reasons.

3. It was known that the variability of the SRI for low-achieving incarcerated youth was consistently quite large across the five measurement occasions. Studies in the future may examine the reliability and validity of the test instrument with item level data.

4. It is also recommended to include more predictors in mixed-effects modeling to further investigate both the inter-individual variability and intra-individual variability in the reading level of low-performing incarcerated youth.
5. With regard to the design factors in power analysis, one might try to vary the size of the variance in the future, while keeping other factors unchanged. This can provide useful information for the influence of the large standard deviation on statistical power.

6. Research on the attitudes of the target population toward reading would also be informative. The motivation of low-performing incarcerated youth is worth investigating in the future. In particular, the reading self-efficacy of these youth is quite helpful in that theoretically (Bandura, 1977, 1997) it may serve as a powerful predictor for their actual reading performance, namely the SRI score.
REFERENCES


Cudeck, R. (2007, Spring). *Analysis of Repeated Measures and Longitudinal Data*. Course notes prepared for Psychology 831.08 lecture at The Ohio State University.


Scholastic, Inc. (2004). *A Study of READ 180 at Shiprock High School in Central Consolidated School District on the Navajo Indian Reservation, New Mexico*. Final report prepared by Scholastic Research and Evaluation Department, New York, NY.


APPENDIX A

MATLAB SCRIPT FOR BOOTSTRAP
Appendix A: A MATLAB Script for Bootstrapping Samples from an Existing Data Set

Continued
%output the sampled data into files
function output2file(fid, data)
    n = size(data,1);
    m = size(data,2);
    for i = 1:n
        for j = 2:6
            if (~isnan(data(i,j)))
                fprintf(fid,'%8d%5d%5d%8d%5d%6d%8.1f%8.1f%5d%5d%6d%5d%5d%5d\n',...
                        i*1000000+data(i,1),...
                        j-2,...
                        power(j-2,2),...
                        data(i,j),...
                        data(i,[7:m])...)
            end
        end
    end
end
APPENDIX B
SAS MACRO
Appendix B: A SAS Macro for Fitting Mixed-Effects Models to a Large Number of Data Sets

%macro run_bootstrap_data;
   %do it = 1 %to 1000;
     /* print to a file */
     proc printto print="C:\SAS_data\MCS_results\mcs_result&it..txt";
     data READ180;
       infile "C:\SAS_data\MCS_data\mcs_data_&it..txt" firstobs=2;
       input Ndys Time Time2 Score TRTGroup Disb Age GrdLvl ReadCat MathCat;
       Timeclass = Time;
     proc mixed data=READ180 method=rem1 asycov asycorr covtest ic;
       class Timeclass;
       model Score = Age ReadCat Disb MathCat*Time GrdLvl*Time2 TRTGroup*Time2
         / noint solution ddfm=satterth covb chisq;
       random intercept Time Time2 /type=un sub=Ndys g gcovr v vcovr solution;
       repeated timeclass / type=simple subject = Ndys r rcorr;
     /* close the file that printto directs to */
     proc printto;
     %end;
%mend run_bootstrap_data;
APPENDIX C

PERL SCRIPT
#!/usr/local/bin/perl
use strict;

use constant tab => "\t";
use constant endl => "\n";

if (@ARGV < 1) {
    print "Usage: extract_results4bootstrap path_to_result_folder\n";
    print "For any questions please contact Jing Zhu at zhu.119@osu.edu\n";
    print "Press enter to continue ...\n";
    <STDIN>;
    exit;
}

my ($rpath) = @ARGV;
my ($FIN, $FVAR, $FFIX, $FPVAL);

die "Can't open a file to write!\n" unless
open FVAR, ">var4rnd.txt" and
open FFIX, ">fixeff.txt" and
open FPVAL, ">pval4fix.txt";
print FVAR "Round	Se	Sb(1,1)	Sb(1,2)	Sb(1,3)	Sb(2,2)	Sb(2,3)	Sb(3,3)\n";
print FFIX "Round	Age	ReadCat	Disb	MathCat*Time	GrdLvl*Time2	Time2*TRTGrp\n";
print FPVAL "Round	Age	ReadCat	Disb	MathCat*Time	GrdLvl*Time2	Time2*TRTGrp\n";

for (my $i = 1; $i < 9999999; ++$i) {
    my $fname = $rpath . "/mcs_result" . $i . ".txt";
    last unless -e $fname;
    die "Can't open a file to read!\n" unless
open FIN, "$fname";
#flag: 1, R matrix, 2, G matrix, 3, fixed effect, 0 else
my $flag = 0;
my $Se = 0;
my @Sb = ();
my @fixeff = ();
my @pval = ();
while (<FIN>) {
    chomp;
    if (/Estimated R Matrix/i) {
        $flag = 1;
    } elsif (/Estimated G Matrix/i) {
        $flag = 2;
    } elsif (/Solution for Fixed Effects/i) {
        $flag = 3;
    }
}
Appendix C Continued.

```csharp
} elsif (/\text{Covariance Matrix for Fixed Effects}/) {
    last;
} elsif ($flag == 1) {
    if (/\^s1\s+(\S+)/) {
        # get Se
        $Se = $1;
        $flag = 0;
    }
} elsif ($flag == 2) {
    if (/\^s1\s+\text{Intercept}\s+\d+\s+(\S+)\s+(\S+)\s+(\S+)/) {
        # get Sb(1,1), (1,2), (1,3)
        $Sb[0] = $1;
        $Sb[1] = $2;
        $Sb[2] = $3;
    } elsif (/\^s2\s+\text{Time}\s+\d+\s+(\S+)\s+(\S+)/) {
        # get Sb(2,2)
        $Sb[3] = $1;
        $Sb[4] = $2;
    } elsif (/\^s3\s+\text{Time2}\s+\d+\s+(\S+)/) {
        # get Sb(3,3)
        $Sb[5] = $1;
        $flag = 0;
    }
} elsif ($flag == 3) {
    if (/\^s\text{Age}\s+(\S+)\s+(\S+)\s+(\S+)\s+(\S+)/) {
        $fixeff[0] = $1;
        $pval[0] = $2;
    } elsif (/\^s\text{ReadCat}\s+(\S+)\s+(\S+)\s+(\S+)/) {
        $fixeff[1] = $1;
        $pval[1] = $2;
    } elsif (/\^s\text{Disb}\s+(\S+)\s+(\S+)\s+(\S+)\s+(\S+)/) {
        $fixeff[2] = $1;
        $pval[2] = $2;
    } elsif (/\^s\text{MathCat}\s*\text{Time}\s+(\S+)\s+(\S+)\s+(\S+)\s+(\S+)/) {
        $fixeff[3] = $1;
        $pval[3] = $2;
    } elsif (/\^s\text{GrdLvl}\s*\text{Time2}\s+(\S+)\s+(\S+)\s+(\S+)\s+(\S+)/) {
        $fixeff[4] = $1;
        $pval[4] = $2;
    } elsif (/\^s\text{Time2}\s*\text{TRTGroup}\s+(\S+)\s+(\S+)\s+(\S+)\s+(\S+)/) {
        $fixeff[5] = $1;
        $pval[5] = $2;
        $flag = 0;
    }
}
```

Continued
for (my $j = 0; $j < @pval; ++$j) {
    $pval[$j] = 0 if ($pval[$j] eq "<.0001");
}
print FVAR $i, tab, $Se, tab, join(tab, @Sb), endl;
print FFIX $i, tab, join(tab, @fixeff), endl;
print FPVAL $i, tab, join(tab, @pval), endl;
close FIN;
}
close FVAR;
close FFIX;
close FPVAL;
APPENDIX D

MATLAB SCRIPT FOR MONTE CARLO
Appendix D: A MATLAB Script for Monte Carlo Simulations

function err_code = MonteCarlo4Read180(pDisb, AgeGrdPool, uRMCatPool4Disb, ...  
    SRMCatPool4Disb, uRMCatPool4Disb, ...  
    SRMCatPool, beta, Sb, se, ...  
    max_t, spl_size, n_round, is_balanced)  

    % Monte Carlo simulation  
    % Usage: error = MonteCarlo4Read180(pDisb, AgeGrdPool, uRMCatPool4Disb, ...  
    %     SRMCatPool4Disb, uRMCatPool, ...  
    %     SRMCatPool, beta, Sb, se, ...  
    %     max_t, spl_size, n_round)  
    % parameters:  
    % pDisb(scale): Bernolli parameter for Disb  
    % AgeGrdPool(inf. x 2): list of students' age and grade level  
    % uRMCatPool4Disb(#age x 2): list of means of Read and Math Cat  
    % SRMCatPool4Disb(#age x 4): list of covariances of Read and Math Cat  
    % for various ages w/ disb  
    % uRMCatPool(#age x 2): list of means of Read and Math Cat  
    % SRMCatPool(#age x 4): list of covariances of Read and Math Cat  
    % for various ages w/o disb  
    % beta: fixed effects of the mixed model  
    % Sb: covariance of random effects of b's  
    % se: standard deviation of random effect of e  
    % max_t: maximum # of time points  
    % spl_size: sampling size for each round, i.e. # of subjects of each sample  
    % default: 1000  
    % n_round: # of rounds, i.e. # of samples to be generated  
    % default: 1  
    %  
    % THE MODEL:  
    % score = beta_0 + beta_1*t + beta_2*t^2 + e  
    % where  
    % beta_0 = beta_00 + beta_01*Age + beta_02*ReadCat + beta_03*Disb + b0  
    % beta_1 = beta_10 + beta_11*MathCat + b1  
    % beta_2 = beta_20 + beta_21*Grd + beta_22*TrtGrp + b2  
    %  
    % fixed effects  
    % -- --  
    % beta = [ 00 01 02 03 10 11 20 21 22 ]  
    % -- --  
    % Sb = Cov([b0,b1,b2])  
    % se = Std(e)  
    %  
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%  
    % To simulate unbalanced data  
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%  
    % Observed portion of subjects with 1, 2, 3, 4 and 5 score measurements  
    % q1: 0.1086  
    % q2: 0.19345
%q3: 0.26042
%q4: 0.20238
%q5: 0.23512

%Observed portion of subjects with score measurement(s) of 1, 2, 3, 4 and 5
%cuml. q1 = SUM(qi)|i=1 to 5: 1 (all subjects have 1st measurement
%cuml. q2 = SUM(qi)|i=2 to 5: 0.89137
%cuml. q3 = SUM(qi)|i=3 to 5: 0.69792
%cuml. q4 = q4+q5: 0.4375
%cuml. q5 = q5: 0.23512
%define p(i) (i=1 to 4) to be the probability that a subject with
% i measurements will continue in the program to receive next
% (i+1 th) measurement
%define p(0) = 1.0 because all subjects have the initial measurement.
%p(0) = 1.0
%p(1) = cuml. q2 / cuml. q1 = 0.89137 / 1.0 = 0.89137
%p(2) = cuml. q3 / cuml. q2 = 0.78297
%p(3) = cuml. q4 / cuml. q3 = 0.62687
%p(4) = cuml. q5 / cuml. q4 = 0.53741
%p(i) v.s. i is linear with fixed intercept of 1.0
%p(i) = -0.1171*i + 1.0
%R2 as high as 0.9938
%This finding will be used to modeling the drop-out pattern of subjects.
%Further, drop-out pattern of subjects has no observal relation with
%other traits of the subjects.
%
%If we define a distribution of # of score mesurements
%its probability mass function (pmf) (qi defined previously) is
%f(x) = (a*x) * (PROD(1-a*i)|i=0 to x-1)
%where a = 0.1171
%its cumularie distribution function (cuml. qi defined previously) is
%F(x) = PROD(1-a*i)|i=0 to x-1
%p(0) ignored
survive_rate = -0.1171*(1:max_t-1) + 1.0;
%default, unbalanced
if (nargin < 12), is_balanced = 0; end
%to avoid odd spl_size
spl_size = round(spl_size/2)*2;
%default # of rounds
if (nargin < 3), n_round = 1; end
%error code: 0, no errors, others, errors
err_code = 0;
%fixed effects
beta_00 = beta(1);
Appendix D Continued

```matlab
beta_01 = beta(2);
beta_02 = beta(3);
beta_03 = beta(4);
beta_10 = beta(5);
beta_11 = beta(6);
beta_20 = beta(7);
beta_21 = beta(8);
beta_22 = beta(9);

%constants
% # of subjects in the list
n_AgeGrdPool = size(AgeGrdPool,1);
% min age of all subjects
minAge = min(AgeGrdPool(:,1));
% max age of all subjects
maxAge = max(AgeGrdPool(:,1));
% a vector of all zeros
zero_vec = zeros(spl_size, 1);
% a vector of all ones
one_vec = ones(spl_size,1);
% mean of [b1,b2,b3] which is assumed to be a zero vector
ub = [0,0,0];

%start timing
tic;
perc_finished = 0;
last_rpt_time = toc;
for i_round = 1:n_round
    % simulate predictors: half Read180, half Comparison
    % TRTGrp
    spl_TrtGrp = zero_vec;
    % first half = 1, rest half 0 (unchanged)
    spl_TrtGrp(1:spl_size/2,1) = 1;
    % Disb
    spl_Disb = binornd(one_vec,pDisb);
    % Age and Grd lvl
    AgeGrdInd = unidrnd(n_AgeGrdPool,spl_size,1);
spl_Age = AgeGrdPool(AgeGrdInd,1);
spl_Grd = AgeGrdPool(AgeGrdInd,2);
    % ReadCat and MathCat
    spl_ReadCat = zero_vec*NaN;
spl_MathCat = spl_ReadCat;
    for age = minAge:maxAge
        iAge = age - minAge + 1;
```

Continued
for disb = 0:1
    isDisbAge = (spl_Disb == disb & spl_Age == age);
    nDisbAge = sum(isDisbAge);
    if (nDisbAge > 0)
        if (disb)
            uRMCat = uRMCatPool4Disb(iAge,:);
            sRMCat = SRMCatPool4Disb(iAge,:);
        else
            uRMCat = uRMCatPool(iAge,:);
            sRMCat = SRMCatPool(iAge,:);
        end
        sRMCat = reshape(sRMCat,2,2);
        curSpl_RMCat = mvnrnd(uRMCat,sRMCat,nDisbAge);
        spl_ReadCat(isDisbAge) = curSpl_RMCat(:,1);
        spl_MathCat(isDisbAge) = curSpl_RMCat(:,2);
    end
end
end
%simulate score for each time point of each sampled subject
%score = beta_0 + beta_1*t + beta_2*t^2 + e
%beta_0 = beta_00 + beta_01*Age + beta_02*ReadCat + beta_03*Disb + b0
%beta_1 = beta_10 + beta_11*MathCat + b1
%beta_2 = beta_20 + beta_21*Grd + beta_22*TrtGrp + b2
%e = N(0,se^2)
%se: scale
%[b0,b1,b2] = N(0,Sb)
%Sb: 3 x 3 matrix

%in matrix notation
%spl_Score = beta_0 + beta_1*Time + beta_2*Time^2 + e
%Where
% beta_0 = beta_00 + beta_01*spl_Age +
%          beta_02*spl_ReadCat + beta_03*spl_Disb + b0
% beta_1 = beta_10 + beta_11*spl_MathCat + b1
% beta_2 = beta_20 + beta_21*spl_Grd + beta_22*spl_TrtGrp + b2

spl_Score = zeros(spl_size, max_t)*NaN;
    b = mvnrnd(ub,Sb, spl_size);
    beta_0 = beta_00 + beta_01*spl_Age + beta_02*spl_ReadCat +
             beta_03*spl_Disb + b(:,1);
    beta_1 = beta_10 + beta_11*spl_MathCat + b(:,2);
    beta_2 = beta_20 + beta_21*spl_Grd + beta_22*spl_TrtGrp + b(:,3);
    for t = 1:max_t
        e = normrnd(zero_vec, se);
        spl_Score(:,t) = beta_0 + (t-1)*beta_1 + power(t-1,2)*beta_2 + e;
    end
\[
\text{%output sampled data into files}
\]  
\text{fname = ['MCS_data/mcs_data_', int2str(i_round), '.txt'];}
\text{fid = fopen(fname,'wt');}
\text{if (fid == -1)}
\text{\hspace{1em}err_code = system('mkdir MCS_data');}
\text{\hspace{1em}if (err_code ~= 0)}
\text{\hspace{2em}disp('Cannot open a file to write!');}
\text{\hspace{2em}return;}
\text{\hspace{1em}fid = fopen(fname,'wt');}
\text{\hspace{1em}if (fid == -1)}
\text{\hspace{2em}err_code = 1;}
\text{\hspace{2em}disp('Cannot open a file to write!');}
\text{\hspace{2em}return;}
\text{end}
\text{end}
\text{fprintf(fid,}
\text{\hspace{1em}'Ndys\tTime\tTime2\tScore\tTRTGrp\tDisb\tAge\tGrdLvl\tReadCat\tMathCat\n');}
\text{for i = 1:spl_size}
\text{\hspace{1em}for j = 1:max_t}
\text{\hspace{2em}only simulate those subject with more than 3 score}
\text{\hspace{2em}measurements, after that survive rate will be used to test}
\text{\hspace{2em}if the subject will stay and next measurement will be taken}
\text{\hspace{2em}if (j > 3 && (~is_balanced))}
\text{\hspace{4em}if (binornd(1,survive_rate(j-1)) == 0), break, end;}
\text{\hspace{1em}fprintf(fid,'\%8d\%5d\%5d\%8.0f\%5d\%5d\%6d\%6d\%8.1f\%8.1f\n', ...}
\text{\hspace{2em}i, ...}
\text{\hspace{2em}j-1, ...}
\text{\hspace{2em}power(j-1,2), ...}
\text{\hspace{2em}spl_Score(i,j), ...}
\text{\hspace{2em}spl_TrtGrp(i), ...}
\text{\hspace{2em}spl_Disb(i), ...}
\text{\hspace{2em}spl_Age(i), ...}
\text{\hspace{2em}spl_Grd(i), ...}
\text{\hspace{2em}spl_ReadCat(i), ...}
\text{\hspace{2em}spl_MathCat(i) ...}
\text{\hspace{1em});}
\text{end}
\text{end}
\text{fclose(fid);}
\text{\%reporting progress}
\text{\%floor(i_round/n_round*100) > perc_finished}
\text{\hspace{1em}perc_finished = floor(i_round/n_round*100);}
\text{\hspace{1em}if (toc - last_rpt_time > 5)}
\text{\hspace{2em}last_rpt_time = toc;}
\]
disp([int2str(perc_finished),'% finished']);
end
end

cpu_time = toc;
t_min = floor(cpu_time/60);
t_sec = round(cpu_time - t_min*60);
disp(['CPU time: ', int2str(t_min), ' min ', int2str(t_sec), ' sec']);