EVALUATING EFFECTIVENESS OF TIER-2 INTERVENTIONS WITHIN A RESPONSE-TO-INTERVENTION FRAMEWORK: A COMPARATIVE ANALYSIS OF CORRECTED MEANS AND PROPENSITY SCORE ANALYSIS METHODOLOGIES

A dissertation submitted to the Kent State University College and Graduate School of Education, Health, and Human Services in partial fulfillment of the requirements for the degree of Doctor of Philosophy

by

Edward Dean Roshong

December, 2009
A dissertation written by

Edward Dean Roshong

B.A., Malone University, 1986

M.A., The University of Akron, 1989

Ph.D., Kent State University, 2009

Approved by

______________________, Co-director, Doctoral Dissertation Committee
Richard Cowan

______________________, Co-director, Doctoral Dissertation Committee
Shawn Fitzgerald

______________________, Member, Doctoral Dissertation Committee
Frank Sansosti

Accepted by

______________________, Interim Director, School of Lifespan Development and
Mary Dellmann-Jenkins Educational Sciences

______________________, Dean, College and Graduate School of
Daniel F. Mahoney Education, Health, and Human Services
This study investigated the effectiveness of a standard protocol Tier 2 reading intervention among third and fifth grade students and the methodologies used to determine the intervention’s effectiveness. Several confounding covariates were observed as a result of utilizing eligibility criteria for assignment to the Tier 2 intervention condition. The biasing effects of these covariates were controlled using traditional ANCOVA and a methodology typically utilized in medical observational studies, propensity score analysis. Although a large amount of research is available on the effectiveness of particular Tier 2 interventions, no research has compared the merits of ANCOVA and propensity score analysis in estimating the effectiveness of these interventions in an applied setting.

Three significant findings were obtained in this study. First, although third grade students receiving Tier 2 reading intervention made significant gains toward closing the grade level achievement gap, their gains were smaller than those of peers receiving only Tier 1 intervention. Among fifth grade students, both groups gained at least one grade level, although there was no difference in gains of students receiving Tier 1 and Tier 2
interventions. Third, similar effect sizes were reported by ANCOVA and propensity score analysis approaches in both the third and fifth grade studies.

Propensity score analysis resulted in similar conclusions while reporting treatment effects in terms of actual criterion scores (i.e., Ohio Achievement Test-Reading). Traditional ANCOVA analysis reported treatment effects as adjusted criterion scores which are not necessarily reflective of achievement test scaling. This study has significant implications for future research and current practice regarding school psychologists’ role in systems consultation, improving achievement for all students, and data-based decision making within a response-to-intervention framework.
ACKNOWLEDGEMENTS

First and foremost, I would like to express appreciation to my dissertation advisor, Dr. Richard Cowan, for his continuous encouragement, editing and most of all his genuine caring during the completion of this project. I would also like to express gratitude to my faculty advisor, Dr. Caven Mcloughlin whose input throughout the doctoral study process has been exceptional. You both helped me clear many hurdles.

I have the deepest appreciation for the sacrifices made by my wife, Janelle, during the completion of this project. You have been just the right kind of encouragement when I needed it. Although there were many family projects that were delayed during these years, I know that what you gave up most was having my full attention when you needed it. I could not have done this without your involvement, encouragement, patience, and prayers. This degree is yours as much as it is mine and I am a better person because of you.

My children, Annie and Lucas hardly remember a time when their dad was not studying or writing. I enjoyed the times we “did our homework together” and am grateful for the times you helped me enter data instead of playing on the family boat. You each sacrificed time with dad as your way of helping me in the process, and you never held it against me. I look forward to making up the time with you. This degree is also yours, and I hope it inspires both of you to aim high in your education.

Finally, I would like to express gratitude to Dr. Jeff Graham and Woodridge Local Schools for your interest and encouragement throughout this process and for allowing me to use school district data records in the completion of the study.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>I INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td>1</td>
</tr>
<tr>
<td>Purpose of the Study</td>
<td>6</td>
</tr>
<tr>
<td>Hypotheses for the Proposed Study</td>
<td>7</td>
</tr>
<tr>
<td>Research Questions</td>
<td>7</td>
</tr>
<tr>
<td>General Hypotheses</td>
<td>8</td>
</tr>
<tr>
<td>Definition of Terms</td>
<td>10</td>
</tr>
<tr>
<td>II REVIEW OF THE LITERATURE</td>
<td>13</td>
</tr>
<tr>
<td>Introduction</td>
<td>13</td>
</tr>
<tr>
<td>Literature Review Related to Legislative Mandates</td>
<td>13</td>
</tr>
<tr>
<td>Overview of Legislative Mandates</td>
<td>13</td>
</tr>
<tr>
<td>Reforms in Student Accountability Standards Mandated by NCLB</td>
<td>14</td>
</tr>
<tr>
<td>Reforms in Teacher Quality Standards Mandated by NCLB</td>
<td>16</td>
</tr>
<tr>
<td>Reforms in Standards for Identifying Scientifically Based Educational Practices in NCLB</td>
<td>17</td>
</tr>
<tr>
<td>Evidence-based Intervention Emphasis in the School Psychology Profession</td>
<td>18</td>
</tr>
<tr>
<td>What Works Clearinghouse</td>
<td>19</td>
</tr>
<tr>
<td>Literature Review Related to a Three-Tier Delivery System</td>
<td>22</td>
</tr>
<tr>
<td>Origins and Development of a Three-Tier Delivery System</td>
<td>22</td>
</tr>
<tr>
<td>Description of Three-Tier Delivery System and Relationship to RTI</td>
<td>23</td>
</tr>
<tr>
<td>Tier 1</td>
<td>24</td>
</tr>
<tr>
<td>Tier 2</td>
<td>27</td>
</tr>
<tr>
<td>Tier 3</td>
<td>29</td>
</tr>
<tr>
<td>Description of Program Evaluation Processes at Each Tier</td>
<td>31</td>
</tr>
<tr>
<td>Literature Review Related to School Psychologist’s Role in Tier 2 Program Evaluation</td>
<td>36</td>
</tr>
<tr>
<td>School Psychologists’ Role in Program Evaluation Suggested by Blueprint Outcome Goals</td>
<td>39</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>School Psychologists’ Role in Program Evaluation Suggested in Blueprint Domains of Competence</td>
<td>41</td>
</tr>
<tr>
<td>Literature Review Related to Evidence-Based Tier 2 Reading Interventions</td>
<td>43</td>
</tr>
<tr>
<td>Selection of Evidence-Based Tier 2 Interventions</td>
<td>44</td>
</tr>
<tr>
<td>Tier 2 Interventions Meeting What Works Clearinghouse Evidence Standards</td>
<td>48</td>
</tr>
<tr>
<td>Additional Evidence-Based Interventions in School Psychology Literature Critical Service Delivery Features</td>
<td>55</td>
</tr>
<tr>
<td>Treatment Fidelity</td>
<td>58</td>
</tr>
<tr>
<td>Instructor-to-Student Ratio</td>
<td>59</td>
</tr>
<tr>
<td>Instructional Term</td>
<td>60</td>
</tr>
<tr>
<td>Instructor Status</td>
<td>61</td>
</tr>
<tr>
<td>Summary of Features Common to Evidence-Based Tier 2 Reading Interventions</td>
<td>62</td>
</tr>
<tr>
<td>Literature Review Related to Propensity Score Analysis Methodology</td>
<td>63</td>
</tr>
<tr>
<td>Selection Bias in Program Evaluations of Tier 2 Intervention Programs</td>
<td>63</td>
</tr>
<tr>
<td>Common Methods of Adjusting for Selection Bias in Evaluations of Tier 2 Intervention Programs</td>
<td>65</td>
</tr>
<tr>
<td>Regression Models</td>
<td>65</td>
</tr>
<tr>
<td>Analysis of Covariance</td>
<td>67</td>
</tr>
<tr>
<td>Weaknesses of Adjusted Means Approaches Such as Regression and ANCOVA</td>
<td>68</td>
</tr>
<tr>
<td>Type VI Error</td>
<td>70</td>
</tr>
<tr>
<td>Propensity Score Analysis: An Alternative Method of Addressing Selection Bias in Program Evaluations of Tier 2 Intervention Programs</td>
<td>73</td>
</tr>
<tr>
<td>Literature Review Related to Applications of Propensity Score Analysis Use of Propensity Score Analysis Methodology in Applied Settings</td>
<td>76</td>
</tr>
<tr>
<td>Early Studies Utilizing Propensity Score Analysis in the General Scientific Literature</td>
<td>77</td>
</tr>
<tr>
<td>Early Applications of Propensity Score Analysis in Mental Health and Social Science Literature</td>
<td>77</td>
</tr>
<tr>
<td>Contemporary Applications of Propensity Score Analysis in the General Scientific Literature</td>
<td>78</td>
</tr>
<tr>
<td>Contemporary Applications of Propensity Score Analysis in the Social Sciences Literature</td>
<td>80</td>
</tr>
<tr>
<td>Contemporary Applications of Propensity Score Analysis in the Criminal Justice Literature</td>
<td>82</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Contemporary Applications of Propensity Score Analysis in the Psychology and Child Development Literature</td>
<td>84</td>
</tr>
<tr>
<td>Contemporary Applications of Propensity Score Analysis in the Substance Abuse Literature</td>
<td>86</td>
</tr>
<tr>
<td>Contemporary Applications of Propensity Score Analysis in the Education Literature</td>
<td>87</td>
</tr>
<tr>
<td>Dual Analyses: Comparisons of Regression Techniques and Propensity Score Analysis in Educational Research</td>
<td>91</td>
</tr>
<tr>
<td>Propensity Score Analysis Applied to Reading Program Evaluation</td>
<td>93</td>
</tr>
<tr>
<td>Gap in the Educational Literature Filled by the Proposed Study</td>
<td>94</td>
</tr>
<tr>
<td>Literature Review Related to Selected Covariates in Propensity Score Analysis and ANCOVA Procedures</td>
<td>97</td>
</tr>
<tr>
<td>Race</td>
<td>97</td>
</tr>
<tr>
<td>Economic Disadvantage</td>
<td>100</td>
</tr>
<tr>
<td>Gender</td>
<td>103</td>
</tr>
<tr>
<td>Disability</td>
<td>105</td>
</tr>
<tr>
<td>Oral Reading Fluency</td>
<td>105</td>
</tr>
<tr>
<td>Summary</td>
<td>108</td>
</tr>
<tr>
<td>III METHODOLOGY</td>
<td>110</td>
</tr>
<tr>
<td>Introduction</td>
<td>110</td>
</tr>
<tr>
<td>Participants</td>
<td>112</td>
</tr>
<tr>
<td>Exclusion of Potential Participants</td>
<td>114</td>
</tr>
<tr>
<td>Instrumentation</td>
<td>116</td>
</tr>
<tr>
<td>DIBELS</td>
<td>116</td>
</tr>
<tr>
<td>IOWA Test of Basic Skills – Reading</td>
<td>117</td>
</tr>
<tr>
<td>Ohio Achievement Test – Reading</td>
<td>118</td>
</tr>
<tr>
<td>Procedure</td>
<td>119</td>
</tr>
<tr>
<td>Assignment of Participants</td>
<td>119</td>
</tr>
<tr>
<td>Intervention Procedures</td>
<td>120</td>
</tr>
<tr>
<td>Data Collection/Coding Procedures</td>
<td>122</td>
</tr>
<tr>
<td>Statistical Analysis</td>
<td>123</td>
</tr>
<tr>
<td>Analysis of Covariance Methodology</td>
<td>124</td>
</tr>
<tr>
<td>Propensity Score Analysis Methodology</td>
<td>125</td>
</tr>
<tr>
<td>Foundations of Propensity Score Analysis</td>
<td>127</td>
</tr>
<tr>
<td>Estimating the Propensity Score</td>
<td>127</td>
</tr>
<tr>
<td>Steps Used to Conduct Propensity Score Analysis</td>
<td>128</td>
</tr>
<tr>
<td>Step 1: Identify and select covariates</td>
<td>128</td>
</tr>
<tr>
<td>Step 2: Assess the initial imbalance in the covariates</td>
<td>129</td>
</tr>
<tr>
<td>Step 3: Calculate the propensity scores</td>
<td>129</td>
</tr>
</tbody>
</table>
Step 4: Stratify the propensity scores into four or five subclassifications. 130
Step 5: Assess the balance on the covariates across the treatment groups 130
Step 6: Estimate and statistically test the difference between the treatment means ........................................ 130
Comparing Treatment Effects Determined by ANCOVA and Propensity Score Analysis Methodologies 132
Hypotheses for the Proposed Study .............................................. 134
Testing the Research Questions .............................................. 134
General Hypotheses .............................................. 137

IV RESULTS ........................................................................ 139
Preliminary Data Analysis .............................................. 139
Demographic Data for Participants .............................................. 139
Descriptive Data .............................................. 140
Analysis of Data .............................................. 142
Research Question 1 .............................................. 142
Research Question 1 Analyzed Using Analysis of Covariance (ANCOVA) .............................................. 143
Hypothesis for Research Question 1 Answered Using Analysis of Covariance (ANCOVA) .............................................. 146
Research Question 1 Analyzed Using Propensity Score Analysis .............................................. 147
Hypothesis for Research Question 1 Answered Using Propensity Score Analysis .............................................. 154
Research Question 2 .............................................. 154
Research Question 2 Analyzed Using Analysis of Covariance (ANCOVA) .............................................. 155
Hypothesis for Research Question 2 Answered Using Analysis of Covariance (ANCOVA) .............................................. 159
Research Question 2 Analyzed Using Propensity Score Analysis .............................................. 160
Hypothesis for Research Question 2 Answered Using Propensity Score Analysis .............................................. 168
Research Question 3 .............................................. 168
Research Question 4 .............................................. 171

V DISCUSSION ........................................................................ 173
Comparison of Results of ANCOVA and Propensity Score Analysis Methodologies .............................................. 174
Third Grade Participants .............................................. 174
Fifth Grade Participants .............................................. 175
Effect Size Comparisons .............................................. 176
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practical Significance</td>
<td>176</td>
</tr>
<tr>
<td>Implications Regarding the Effectiveness of the Intervention</td>
<td>177</td>
</tr>
<tr>
<td>Implications Regarding the Appropriateness of the Methodology Used in Analysis of the Data</td>
<td>178</td>
</tr>
<tr>
<td>Delimitations of the Study</td>
<td>181</td>
</tr>
<tr>
<td>Quasi-Experimental Design</td>
<td>181</td>
</tr>
<tr>
<td>Exclusion of Potential Participants</td>
<td>182</td>
</tr>
<tr>
<td>Limitations of the Study</td>
<td>184</td>
</tr>
<tr>
<td>Limitations with Respect to Treatment Fidelity</td>
<td>184</td>
</tr>
<tr>
<td>Limitations with Respect to Instrumentation</td>
<td>186</td>
</tr>
<tr>
<td>DIBELS</td>
<td>186</td>
</tr>
<tr>
<td>Ohio Achievement Test – Reading (OAT-R)</td>
<td>187</td>
</tr>
<tr>
<td>Limitations Concerning Reliability and Standard Error of Measurement</td>
<td>188</td>
</tr>
<tr>
<td>Limitations Concerning Scaled Score Confidence Intervals</td>
<td>189</td>
</tr>
<tr>
<td>Limitations Concerning Sensitivity of Instrumentation</td>
<td>190</td>
</tr>
<tr>
<td>Limitations Concerning the Timing of Test Administrations</td>
<td>190</td>
</tr>
<tr>
<td>Limitations Concerning Content Validity of the Dependent Variable</td>
<td>191</td>
</tr>
<tr>
<td>Potential Multicollinearity of Variables</td>
<td>192</td>
</tr>
<tr>
<td>Limitation Related to Sample Size</td>
<td>194</td>
</tr>
<tr>
<td>Implications for the Profession of School Psychology</td>
<td>196</td>
</tr>
<tr>
<td>Implications for Future Research</td>
<td>198</td>
</tr>
</tbody>
</table>

REFERENCES ........................................................................................................ 201
LIST OF TABLES

Table                          Page
1.  Data-Based Problem-Solving Model…………………………………………………30
2.  Competencies of School Psychologists Related to Closing Achievement Gaps……..38
3.  Targeted Interventions Meeting What Works Clearinghouse Rigorous Evidence Criteria………………………………………………………………………………..46
5.  Grades 4 and 8 Mean Reading Scaled Scores by Student Eligibility for Free/reduced Lunch for Available Years 1998-2005…………………………………………102
6.  Grades 4 and 8 Mean Reading Scaled Scores by Gender for Available Years 1998-2005……………………………………………………………………………….104
7.  Participants Before and After Exclusion Procedures………………………………115
8.  Descriptive Statistics for Third Grade Participants………………………………140
9.  Descriptive Statistics for Fifth Grade Participants………………………………141
10. Correlation Matrix for All Variables – Third Grade………………………………144
11. Results of the ANCOVA Analysis of Grade 3 Reading Gain Scores Using MLR Model 4……………………………………………………………………………….145
12. Descriptive Data for Third Grade Reading Achievement Gains…………………146
13. Descriptive Statistics for the Criterion Variable and Covariates – Grade 3………148
14. Comparison of Differences Between Tier 1 and Tier 2 Intervention Groups on Covariates Before and After Propensity Group Formation – Grade 3………………149
15. Results for the Logistic Regression Model – Grade 3……………………………..150
16. Estimated Treatment Effects on Third Grade Reading Gain Scores Using Propensity Score Groups……………………………………………………………………153
17. Correlation Matrix for All Variables – Grade 5……………………………………156
18. Results of the ANCOVA Analysis of Grade 5 Reading Gain Scores
   Using MLR Model 6………………………………………………………………………158
19. Descriptive Data for Grade 5 Reading Achievement Gains………………………159
20. Descriptive Statistics for the Criterion Variable and Covariates – Grade 5………161
21. Comparison of Differences Between Tier 1 and Tier 2 Intervention Groups
    on Covariates Before and After Propensity Group Formation – Grade 5………….163
22. Results for the Logistic Regression Model – Grade 5……………………………164
23. Estimated Treatment Effects on Grade 5 Reading Gain Scores Using
    Propensity Score Groups……………………………………………………………..167
CHAPTER I
INTRODUCTION
Statement of the Problem

The No Child Left Behind Act of 2001 (NCLB) made numerous requirements to increase the accountability of public schools. For example, schools must annually assess reading and mathematics achievement of students in grades 3-8. They must also make adequate yearly progress (AYP) in the general student body and within each subgroup of students. Achievement goals must be achieved by all subgroups of students based upon race, ethnicity, economic disadvantage, disability, and limited English proficiency. Research-based interventions must be implemented for all students who do not meet the assessment benchmarks. In addition, schools that do not meet the adequate yearly progress targets of NCLB face multiple consequences such as administrative restructuring, incurring the costs of providing supplemental educational services, and the costs of transporting students to more successful neighboring districts.

Because NCLB focuses on the achievement of every student, closing achievement gaps between the general population and Adequate Yearly Progress (AYP) subgroups have been of interest to lawmakers and educational researchers alike. Of particular interest is the goal of closing the achievement gap between the general population and specific groups of students including African-American students, students with disabilities, and English language learners (State Board of Education Closing Achievement Gaps Task Force, 2003). It is of critical importance to identify those
interventions that are successful in helping schools close these achievement gaps through the most effective research methods.

This key legislative initiative is having an impact on the role and function of school psychologists. In the next decade, school psychologists are envisioned playing an increasing role in addressing the academic achievement of students with and without disabilities in the general education environment (Burns, 2007; Burns et al., 2006; Ysseldyke et al., 2006). *The Blueprint for Training and Practice III* (Ysseldyke et al., 2006), a guiding document in the profession of school psychology, identifies those functional and behavioral competencies within the role and function of school psychologists as they are engaged in practice at multiple levels in the intervention process. *Blueprint III* suggests that school psychologists should be working towards (a) improving competencies for all students, and (b) building and maintaining the capacities of systems to meet the needs of all students. Specific functional competencies promoted by *Blueprint III* that are relevant to the task of closing the achievement gap include enhancing the development of cognitive and academic skills in all students, systems-based service delivery, and data-based decision making and accountability.

The delivery system envisioned to achieve these goals has a long history in the public health and prevention literature (Ysseldyke et al., 2006). Interventions are delivered to students in three levels: universal, targeted, and intensive to accomplish the goal of closing achievement gaps and meeting adequate yearly progress goals (Marston, Muyskens, Lau, & Canter, 2003). The first tier, universal intervention, is designed to meet the needs of approximately 85% of learners, and involves a challenging research
based general curriculum. The second tier, targeted, represents interventions for 10-20% of students who have not met achievement benchmarks and who require a more intensive level of intervention. The third tier, intensive, represents approximately 5% of students who have not adequately responded to Tier 2 targeted interventions, and who therefore require highly individualized and tailored interventions, including students with disabilities (Burns, Deno, & Jimerson, 2007).

Measurement of student progress toward meeting AYP requirements in reading and mathematics is a process ultimately determined by each state’s department of education (NCLB, 2001). Federal law mandates only that the annual assessments are criterion-based and measure progress toward meeting state-determined curriculum standards. Other measures of student progress such as curriculum-based measurement procedures are available to researchers as measures of intervention effectiveness (Deno, 1985; Fuchs & Fuchs, 2002; Shinn, 1989, 2002). In particular, measures of oral reading fluency have been demonstrated as sensitive to small changes in skill acquisition and efficient in administration to be used in progress monitoring (Baker et al., 2008; Chidsey-Brown & Steege, 2005; Shinn, 1998). Nevertheless, because the standards-based, state designed assessments are the basis of AYP determinations, evaluation of intervention programs use scores on these high stakes state assessments as the criterion variable in classifying interventions as effective.

Program evaluation at each tier of this model typically involves different research designs and evaluation methods. At Tier 1, large scale experimental designs and descriptive research are appropriate because all students are initially assigned to the
interventions. Evaluating the progress of students receiving the most intensive individualized interventions at Tier 3 is most appropriately conducted using single subject research designs (Walcott & Riley-Tillman, 2007). Program evaluation of Tier 2 interventions typically presents a research challenge for researchers and practitioners alike. Between-group analysis is the preferred method of evaluating the effectiveness of Tier 2 interventions because they are delivered in a group format and involve larger numbers of students (Burns et al., 2006). Because students are assigned to these interventions as a result of poor response to the general curriculum, the sample of students participating in Tier 2 interventions is inherently biased and disproportionately representative of ethnic minorities, English language learners, students with disabilities, and economically disadvantaged students (Newell & Kratochwill, 2007).

Title 1 of the No Child Left Behind Act of 2001 (NCLB) provides funding for local school districts with significant economically disadvantaged populations to establish targeted intervention programs for students with low achievement scores in reading and mathematics. A body of research exists identifying the inherent sample biases among students assigned to Tier 2 interventions such as Title 1 programs (McMaster, Fuchs, Fuchs, & Compton, 2005). For example, students participating in Tier 2 interventions such as Title 1 reading programs are assigned specifically because they have low reading test scores. Assigning students to a treatment group based upon reading achievement test scores would create a biased treatment group of poor readers (U.S. Department of Education Institute of Education Sciences, 2003). Research has also shown that racial minorities, economically disadvantaged students, and English language
learners have historically been over represented in Title 1 reading intervention programs (Carlisle, Schilling, Zeng, Cortina, & Kleyman, 2006). These factors also represent sources of inherent bias in the samples under study.

The presence of a known systematic bias presents a critical challenge to school psychologists acting in the role of intervention program evaluators. The most common methods of controlling for this bias are the use of systematic regression techniques or Analysis of Covariance (ANCOVA). While this creates a method of controlling for the inherent selection bias, it also creates a new type of error called type VI error (Newman, Fraas, Newman, et al., 2005). A Type VI error is a mismatch between the research question and the selected method of analysis. If the original challenge was to identify relatively large scale interventions that could successfully close the gap between general students and low achieving students, African-American students, or economically disadvantaged students, the method of controlling for the systematic bias removes the power of the researcher to answer the original research question. The ANCOVA technique allows the researcher to compare the gains of groups by adjusting the means to control for the selection bias. However, the original aim of the researcher is to intentionally target students in a biasing condition and compare their unadjusted means.

Researchers in the medical field have used an analytic technique called propensity score analysis (D’Agostino, 1998; Rosenbaum, 2002; Rosenbaum & Rubin, 1983, 1984) as an alternative to regression or ANCOVA techniques to evaluate outcomes in situations where there is an inherently biased sample. For example, when it has been impossible to randomly assign patients to the smoking or non-smoking group or the cancer or non-
cancer group, this technique has been shown effective in helping the research answer the question of whether a specific drug is effective in treating cancer in patients who are smokers.

The propensity score analysis technique can be applied to educational research using the same rationale. If prior student achievement is the basis of assignment to a Tier 2 targeted intervention, it is desirable to determine if the interventions are significantly effective without systematically treating analyses as if students are entering without an educational deficit. School psychologists evaluating these programs wish to know if the intervention works particularly when the students are known to be low achievers.

**Purpose of the Study**

The purpose of this study was to investigate the methods used to evaluate the effectiveness of Tier 2 interventions. Specifically, the reading achievement test results of students participating in a standard protocol reading intervention program were analyzed using two methods: Analysis of Covariance and Propensity Score Analysis. Because school psychologists play an increasingly pivotal role in improving the achievement of all learners and increasing their influence in designing and implementing interventions within the general curriculum, they must utilize the best methods to evaluate those programs. The problem of selection bias is inherent in any evaluation of a program targeted to an AYP subgroup as mandated by NCLB. Propensity score analysis, previously used in applications to medical research, offers an alternative method of analyzing data in observational studies with the additional benefit of retaining the bias in the original data. Although a large amount of research is available on the use of methods
employing adjusted means to this problem, no research has compared the use of adjusted means methods to propensity score analysis on a set of data from a typical Tier 2 intervention in a local school system.

Hypotheses for the Proposed Study

The current study proposed to address this gap in the literature by conducting a local program evaluation of a Tier 2 reading intervention program in which the unit of study is at the student level. Propensity score stratification, which utilizes more of the available data than propensity score matching, and can better account for the differences in group sizes between treatment and control groups will be utilized to balance subjects on known covariates. The research questions were addressed by treating the data both by Analysis of Covariance (ANCOVA) and propensity score stratification, and a comparison of the research conclusions will be made. The value of the study was to illustrate whether there is ample evidence that school psychologists should utilize propensity score analysis rather than traditional ANCOVA to conduct local program evaluations of Tier 2 intervention programs.

Research Questions

The following research questions were addressed in this study:

1. Will third grade students who received the Tier 2 standard protocol reading intervention make different gains in reading achievement than students who received only the Tier 1 reading intervention when considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race?
2. Will fifth grade students who received the Tier 2 standard protocol reading intervention make different gains in reading achievement than students who received only the Tier 1 reading intervention when considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race?

3. Will there be a significant difference in the effect sizes reported from a program evaluation of the effectiveness of the Tier 2 reading intervention among third grade students when the analysis is conducted using traditional Analysis of Covariance (ANCOVA) compared to propensity score analysis?

4. Will there be a significant difference in the effect sizes reported from a program evaluation of the effectiveness of the Tier 2 reading intervention among fifth grade students when the analysis is conducted using traditional Analysis of Covariance (ANCOVA) compared to propensity score analysis?

General Hypotheses

Hypotheses for Research Question 1: Third grade students who received the Tier 2 standard protocol reading intervention will make different adjusted achievement gains on the Ohio Achievement Test in reading (p<.05) than students who received only the Tier 1 reading intervention after considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race. Third grade students who received the Tier 2 standard protocol reading intervention will make different achievement gains on the Ohio Achievement Test in reading (p<.05) than students who received only the Tier 1 reading intervention when controlling for prior reading achievement, oral reading fluency, economic disadvantage, and race.
Hypotheses for Research Question 2: Fifth grade students who received the Tier 2 standard protocol reading intervention will make different adjusted achievement gains on the Ohio Achievement Test in reading (p<.05) than students who received only the Tier 1 reading intervention after considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race. Fifth grade students who received the Tier 2 standard protocol reading intervention will make different achievement gains on the Ohio Achievement Test in reading (p<.05) than students who received only the Tier 1 reading intervention when controlling for prior reading achievement, oral reading fluency, economic disadvantage, and race.

Hypothesis for Research Question 3: There will be a significant difference (p<.05) between the effect sizes estimating the difference between Ohio Achievement Test Reading gains for third grade students who participated in the Tier 2 intervention compared to students who received only the Tier 1 reading intervention when the differences are estimated using ANCOVA and propensity score analysis.

Hypothesis for Research Question 4: There will be a significant difference (p<.05) between the effect sizes estimating the difference between Ohio Achievement Test Reading gains for fifth grade students who participated in the Tier 2 intervention compared to students who received only the Tier 1 reading intervention when the differences are estimated using Analysis of Covariance (ANCOVA) and propensity score analysis.
Definition of Terms

1. Achievement gap – In this study the term refers to the difference in mean achievement between the majority population and that of a demographic subgroup.

2. Adjusted means comparisons – In this study the term means methods of statistically controlling for selection bias when experimental and control groups are not randomly assigned. Analysis of Covariance (ANCOVA) or multiple regression analysis are utilized to statistically estimate the posttest means to what they would have been if the groups had started out equally on the covariates (i.e. the state of equalization of group characteristics when random assignment is possible).

3. Adequate Yearly Progress (AYP) – In this study the term refers to a measurement of student achievement mandated by the No Child Left Behind Act. Every public school district must demonstrate achievement of annual targets in standardized tests of reading and mathematics for its students in grades K-12. All students must achieve proficiency in reading and mathematics by the year 2014.

4. AYP Subgroups – In this study the term means the demographic subgroups of students each district must demonstrate are making adequate yearly progress. The targeted subgroups are economically disadvantaged students, students with disabilities, students with limited English proficiency, and students in five racial minority groups: African-American, Asian, Hispanic, Multiracial, and Native American.

5. Curriculum-based measurement – In this study the term means a set of standardized and validated short duration tests that are used by special education and general education teachers for the purpose of evaluating the effects of their instructional
programs in the areas of basic reading skills, mathematics computation, spelling and written expression.

6. Observational studies – In this study means an empiric investigation of the effects caused by a treatment, policy, or intervention in which it is not possible to assign subjects at random to treatment or control, as would be done in a controlled experiment.

7. Problem solving model – In this study means a series of process steps designed to assist a team of educators in making intervention decisions when attempting to move a student from a current state to a future desired state.

8. Propensity score analysis – In this study means an analytic method that uses propensity scores to adjust the comparison of non-randomized group means for selection bias due to systematic differences on a set of covariates.

9. Response to intervention – In this study means the systematic use of assessment data to allocate instructional resources efficiently within a three-tiered model of intervention in order to enhance learning outcomes.

10. Scientifically based research – In this study the term means instructional methods that have been shown to produce reliably positive achievement results and have demonstrated those results through rigorous research practices.

11. Standard protocol intervention – In this study the term means practices of providing academic interventions that are delivered in accord with a predetermined plan, often in small groups, and supported by rigorous empirical evidence.

12. Tier 2 interventions – In this study the term means interventions targeted at approximately 15-20% of students who have not met achievement benchmarks through
exposure to universal intervention and who require a more intensive level of intervention, typically a standard protocol intervention.

13. Title 1 – In this study the term means federal funding under the No Child Left Behind act targeted at closing the achievement gap for economically disadvantaged students. Funds may be used for a variety of supplemental intervention programs in reading and mathematics that are supported by scientifically based research.

14 Treatment integrity – In this study the term means the extent to which an intervention is implemented as intended.

15. Type VI error – In this study the term means a mismatch between the research question and the selected method of analysis.
CHAPTER II

REVIEW OF LITERATURE

Introduction

This chapter is designed to both (a) review a body of literature to support the need for the present study, and (b) provide a rationale for the method of analysis used in the study. The chapter begins with a general review of the literature on the statutory mandates that contribute to the need to develop improved analytic methods in evaluating Tier 2 interventions. The origin and application of a three tiered delivery system is then described. The influence these factors have on the role of the school psychologist in the intervention process is then examined. Next, a brief review of effective Tier 2 reading interventions is presented. Traditional methods of evaluating the effectiveness of Tier 2 interventions using adjusted-means methods is described, including the pitfalls associated with this methodology. Propensity score analysis, its role in studying data obtained from inherently biased samples, and its advantages over adjusted-means methods is described. The chapter concludes with a review of the literature establishing the rationale for selecting particular covariates for the analysis and proposed analysis for this study.

Literature Review Related to Legislative Mandates

Overview of Legislative Mandates

The No Child Left Behind Act of 2001 (NCLB) was the most comprehensive legislative effort aimed at improving academic achievement in elementary and secondary
schools in the United States in the past forty years (Simpson, LaCava, & Graner, 2004). This legislation implements reforms in student accountability standards, teacher quality standards, standards for identifying scientifically-based educational practices, and increasing school districts’ control over federal funds by introducing more flexibility in using funds provided through the Act.

Because NCLB focuses on the achievement of every student, closing achievement gaps between the general population and Adequate Yearly Progress (AYP) subgroups have been of interest to lawmakers and educational researchers alike. Of particular interest is the goal of closing the achievement gap between the general population and specific groups of students including African-American students, students with disabilities, and English language learners (State Board of Education Closing Achievement Gaps Task Force, 2003). It is of critical importance to identify those interventions that are successful in helping schools close these achievement gaps through the most effective research methods.

*Reforms in Student Accountability Standards Mandated by NCLB*

No Child Left Behind requires that by the end of the 2013-2014 school year, all students in US schools must be proficient in mathematics and reading. Individual schools, districts and states must meet this standard for every student, including students not previously required to pass accountability standards (e.g., students with disabilities and students with limited English proficiency). In order to determine if all students are proficient, NCLB requires states to implement annual standardized assessment of students in mathematics and reading in grades 3-10. Adequate yearly progress (AYP)
benchmarks must be met for the general population of students as well as for students in several underperforming subgroups previously identified with an “achievement gap” when compared to the general student population. The targeted subgroups are economically disadvantaged students, students with disabilities, students with limited English proficiency, and students in five racial minority groups: African-American, Asian, Hispanic, Multiracial, and Native American (State Board of Education Closing Achievement Gaps Task Force, 2003).

When individual schools and school districts are successful in meeting AYP benchmarks for every subgroup, they often receive public recognition for the high achievement of their students. Individual teachers and school administrators in these schools are also eligible to receive positive recognition in the form of awards and acknowledgment at the local and state level. The federal law does not initiate any positive recognition or financial incentive for districts, administrators, or teachers at the federal level.

Individual schools and school districts that do not meet the AYP benchmarks for two years in a row either school wide or in any of the targeted subgroups are subject to being designated as “needs improvement” (NCLB, 2001). These schools must offer parents the opportunity to allow their children attending low-performing schools to attend a different, high-performing public school either within the same school district or in a neighboring school district. Districts that fail to meet AYP for 3 years in a row must offer parents supplemental educational services for their children. Supplemental educational services are private tutoring sessions provided at public expense by diverting Title 1
funds to private tutoring service organizations. Other sanctions that could result from failure to meet AYP over multiple years include replacing the building principal, state takeover of the school, or closing of the school entirely (NCLB, 2001).

Reforms in Teacher Quality Standards Mandated by NCLB

Interventions that are designed to improve the reading achievement of students who are not responding to the general curriculum must be implemented by highly skilled staff. The reading achievement of students is influenced by many factors, such as the quality of the schools they attend, the quality of their early language experiences, family income and educational level, instructional methods used in their schools, and the quality of their teachers (Ehrenberg & Brewer, 1994). A key provision in the accountability standards of NCLB was a direct result of a substantial body of research demonstrating a link between teacher quality and the academic results in the achievement of their students (Darling-Hammond, 1999; Darling-Hammond & Youngs, 2002; Ehrenberg & Brewer, 1994). The legislation includes the requirement that all teachers in Title 1 schools should have been “highly qualified” by the end of the 2005-2006 school year (NCLB, 2001). Although individual states may determine the exact minimum standards to meet the definition of highly qualified, the federal legislation creates some mandatory provisions. In order to meet the highly qualified standard, teachers must possess expertise in the core content area of their teaching and must have specific pedagogical skills in teaching that content area (U.S. Department of Education Office of the Under Secretary, 2003).

The mandatory provisions for highly qualified teachers include the following:

1. Must hold a bachelor’s degree
2. Must demonstrate content area expertise in the content area of their teaching
3. Teachers of core academics should have been highly qualified by 2005-2006
4. New elementary teachers must meet state testing criteria for content knowledge and teaching skills
5. New teachers at the middle and high school level must have an academic major in each content area they teach, a masters degree in the content area, or pass a state test in the content area
6. Teachers already in the field must meet the same standards as new teachers or demonstrate competency as each state defines (U.S. Department of Education Office of the Under Secretary, 2003).

Reforms in Standards for Identifying Scientifically Based Educational Practices in NCLB

The term scientifically based research (SBR) occurs more than 100 times in NCLB (Simpson et al., 2004). The Act requires recipients of federal Title 1 funds to implement educational practices that have been determined to be effective based on SBR. Title 1 is federal funding under the No Child Left Behind act targeted at closing the achievement gap for economically disadvantaged students. Funds may be used for a variety of supplemental intervention programs in reading and mathematics that are supported by scientifically based research. Although the disciplines of education and psychology agree that interventions should be validated using rigorous research standards, there is no universal set of criteria for identifying scientifically based research at this time. Within the field of school psychology, the term scientifically based research refers to instructional methods that have been shown to produce reliably positive
achievement results and have demonstrated those results through rigorous research practices (Algozzine, 2003). These practices are preferably subject to rigorous standards of peer-reviewed research (U.S. Department of Education Institute of Education Sciences, 2003).

Evidence-based Intervention Emphasis in the School Psychology Profession

The profession of school psychology has been instrumental in the development and implementation of evidence-based interventions (EBIs) as well as standards to identify those interventions that produce reliable results. Researchers have identified critical research agendas for the field that include focusing the profession’s research in the domains of diagnosis/classification and assessment in school psychology, consultation, prevention and intervention, program evaluation and research and evaluation of practice (Kratochwill & Stoiber, 2000). The profession’s leaders have also recommended that research initiatives in the field be reframed in congruence with national school change initiatives, including emphasis on identifying interventions that produce measurable results (Kratochwill & Shernoff, 2004; Kratochwill & Stoiber, 2000; Kratochwill, Stoiber, & Gutkin, 2000). The need for such a focused effort is underscored by surveys of school psychology trainers indicating that relatively low percentages of trainers were familiar with specific EBIs (Shernoff, Kratochwill, & Stoiber, 2003). In order to improve the emphasis on EBIs, training programs need to adopt the scientist-practitioner model and emphasize training school psychologists to implement standard protocol interventions (Kratochwill & Shernoff, 2004).
In order to address these priorities, the Task Force on Evidence-Based Interventions in School Psychology was established by APA Division 16 to identify, code, and review studies of interventions for school-aged children and their families (Shernoff et al., 2003). The manual describes specific criteria for rating intervention studies on a four point scale specifying the level of evidence (i.e., 3=strong evidence/support, 2=promising evidence/support, 1=marginal or weak evidence/support, 0=no evidence/support). The coding manual describes eight key features on which the research studies are rated, including: (a) reliability and validity of measurement, (b) existence of a comparison group, (c) statistically significant outcomes, (d) educational/clinical significance, (e) explicitly identifiable intervention components, (f) implementation fidelity, (g) replication, (h) site of implementation, and (i) follow-up assessment (Task Force on Evidence-Based Intervention in School Psychology, 2003). Implementation of these criteria will allow for dissemination of useful information on the effectiveness of interventions to practicing school psychologists and school psychology trainers, with the ultimate goal of closing the gap between research and practice (Shernoff et al., 2003).

**What Works Clearinghouse**

In July 2004 the U.S. Department of Education launched a website dedicated to disseminating evaluative information about commercially available curricula and programs to all professional educators and policymakers (Viadero, 2004). The *What Works Clearinghouse* was developed over a two year period at a cost $18.5 million and
was designed to be a centralized source of information on the extent to which the claims of available strategies and programs met the standards of scientifically based research. In order to identify scientifically based research practices, the What Works Clearinghouse utilizes a Design and Implementation Assessment Device instrument and a related protocol to determine the extent to which the methods used in reviewed research are supported by scientific evidence (U.S. Department of Education Institute of Education Sciences, 2007).

School psychologists working within a three-tier integrated systems model of interventions can apply the data reviewed at What Works Clearinghouse in all three tiers. The intended application of each intervention is categorized in multiple ways to allow the practitioner to determine the appropriateness of implementation. For example, beginning reading interventions are classified as targeted to fluency, alphabetics, comprehension, or general reading achievement. In addition, the target population and intended setting for each intervention is identified in a program review (U.S. Department of Education Institute of Education Sciences, 2008). Interventions may be classified suitable for core programs (Tier 1), supplemental programs (Tier 2), or intensive tutoring (Tier 3).

The Design and Implementation assessment Device and the related protocol have been criticized by many educators and educational researchers. The primary reason for this controversy is that randomized experimental group design methodology is the preferred strategy for demonstrating scientific evidence. Randomized experimental control group designs are rarely used in educational research (U.S. Department of Education Institute of Education Sciences, 2003). However, although there is
controversy regarding the particular method of determining scientific evidence, scholars agree that using effective methods benefits all students, including students with disabilities (Algozzine, 2003; Shavelson & Towne, 2002). School psychologists have contributed significantly to this discussion by developing rigorous standards for identifying interventions that produce reliable results (Kratochwill & Shernoff, 2004; Kratochwill & Stoiber, 2000; Kratochwill et al., 2000), setting prioritized research agendas for the profession (Kratochwill & Stoiber, 2000), and establishing a national task force to coordinate its efforts (Task Force on Evidence-Based Intervention in School Psychology, 2003).

The review process and application of the review protocol at What Works Clearinghouse produces a set of ratings for each intervention in the following domains: (a) improvement index, (b) evidence rating, and (c) extent of evidence. The improvement index represents the difference between the mean percentile ranks of students in the intervention group compared to the mean percentile rank of students in the comparison group in terms of their relative levels of performance on the outcome variable. Interventions reviewed by the Clearinghouse are ranked overall based on the improvement index on values ranging from -50 to +50. The evidence rating considers the factors of quality of research design, statistical significance of between-group differences, the magnitude of between-group differences, and the consistency of replication of results across studies. The extent of evidence refers to the number of studies and the number of students who participated in those studies. Interventions are classified as having either
small or medium to large extent of evidence (U.S. Department of Education Institute of Education Sciences, 2009).

Literature Review Related to a Three-Tier Delivery System

*Origins and Development of a Three-Tier Delivery System*

Although there is no federal mandate to utilize a particular delivery system to achieve the goals of NCLB and the Individuals with Disabilities Education Act (IDEA, 2004), the majority of states have endorsed some form of three-tiered delivery system to address the various levels of student response to interventions (Brown-Chidsey & Steege, 2005; Griffiths, Parson, Burns, VanDerHeyden, & Tilly, 2007). Researchers investigating the best practices in applying response to intervention (RTI) technology overwhelmingly advocate a three-tiered delivery system (Burns & Coolong-Chaffin, 2006; Burns et al., 2007; Fuchs & Fuchs, 2007; Marston et al., 2003; Sugai & Horner, 2006; Vaughn & Roberts, 2007). The delivery system envisioned to achieve these goals has a long history in the public health and prevention literature (Sugai & Horner, 2006; Ysseldyke et al., 2006). The concept of a tiered delivery system, though optimized in IDEA 2004, began as a general education initiative. It was first introduced as a system for strengthening the achievement of students in the general education setting and working in the general curriculum (Griffiths et al., 2007). Response to intervention can be conceptualized as the systematic use of assessment data to allocate instructional resources efficiently within a three-tiered model of intervention in order to enhance learning outcomes (Burns & VanDerHeyden, 2006).
Description of Three-Tier Delivery System and Relationship to RTI

The main structural feature of the tiered intervention model is a delivery of increasingly intensive interventions to a decreasing pool of students requiring the interventions to demonstrate adequate instructional or behavioral progress (Fuchs & Fuchs, 2007). Interventions are delivered to students in three levels (i.e., universal, targeted, and intensive) to accomplish the goal of closing achievement gaps and meeting adequate yearly progress goals (Marston et al., 2003). This delivery system has been graphically represented in a pyramid shape, with universal interventions comprising the base, targeted interventions comprising the middle, and intensive interventions comprising the apex of the pyramid (Brown-Chidsey & Steege, 2005; Burns & Coolong-Chaffin, 2006; Sugai & Horner, 2006; Ysseldyke et al., 2006) (see Figure 1).
Tier 1

The first tier, universal intervention, is designed to meet the needs of approximately 85% of learners, and involves a challenging research-based general curriculum (Vaughn & Roberts, 2007). This tier of intervention takes place for all students in the general education environment and is delivered by general education teachers. Assessment of key progress benchmarks is conducted at least three times per academic year to determine if students are responding appropriately to this level of intervention. The assessment system is typically a curriculum-based measure that is brief.
yet has strong predictive validity (Fuchs & Fuchs, 1999; Shinn, 1989). Curriculum-based measurement is “a set of standardized and validated short duration tests that are used by special education and general education teachers for the purpose of evaluating the effects of their instructional programs in the areas of basic reading skills, mathematics computation, spelling and written expression” (Shinn, 2002, p. 671). The Dynamic Indicators of Basic Early Literacy (DIBELS; Kaminski & Good, 1996) is a commonly utilized benchmark assessment, although many other options are available (Brown-Chidsey & Steege, 2005; Burns & Coolong-Chaffin, 2006).

The DIBELS reading assessments primarily measure prereading skills at grades K-1 and oral reading fluency skills at grades 1-6 (Brown-Chidsey & Steege, 2005). The DIBELS assessments include measures of initial sound fluency (ability to recognize and produce the initial sound in an orally presented word) at grade K; phoneme segmentation fluency (ability to segment three- and four-phoneme words into their individual phonemes) at grades K-1; nonsense word fluency (ability to orally read non-words) at grades K-2; and oral reading fluency in grades 1-6 (Jenkins, Hudson, & Johnson, 2007; University of Oregon Center on Teaching and Learning, 2009). Measuring oral reading fluency involves listening to the student read a selected passage from grade-level reading material for a one minute interval and counting the number of words read correctly. Because the progress of all students is monitored at least three times per year (i.e., fall, winter, spring), any students who demonstrate a poor response to universal intervention may be identified for inclusion in Tier 2 intervention activities. Poor response to the universal intervention is typically operationalized as failure to meet the benchmark
criteria at grade level identified by the curriculum-based measure (Burns & Coolong-Chaffin, 2006; Jenkins et al., 2007). The benchmark criteria are based on both national normative criteria and local norms for the average performance of students in the same grade at the child’s school (Brown-Chidsey & Steege, 2005).

School psychologists can play several important roles in the activities at Tier 1. Primarily, they offer important contributions as consultants with school administrators regarding the assessment system. A valid system of data collection to effectively monitor student progress is fundamental to the success of RTI (Burns, Dean, & Klar, 2004; Fuchs, 2003; Gresham, 2002). School psychologists have expertise in the assessment process and should assist teachers and administrators in choosing assessment tools that are reliable and valid. School psychologists also should play an important role as consultants to administrators and teachers in the selection of the universal reading intervention that is the core curriculum by utilizing their skills in identifying curricula that conform to the standards of evidence-based-research (Burns and Coolong-Chaffin, 2006; Kratochwill & Shernoff, 2004; Kratochwill & Stoiber, 2000; Kratochwill et al., 2000). As individual student performance is assessed and compared to both national and local benchmarks, school psychologists are also invaluable members of a team interpreting this data and determining which students need more intensive intervention. Students scoring below specific cut scores on achievement benchmarks are assigned a more frequent and intensive level of intervention (Brown-Chidsey & Steege, 2005; Burns et al., 2007).
Tier 2

The second tier, targeted intervention, represents interventions for students who have not met achievement benchmarks and who require a more intensive level of intervention (Burns et al., 2007; Sugai & Horner, 2006). Despite the utilization of a high quality core curriculum, 15-20% of the student population may not make sufficient progress to meet accountability standards (Griffiths et al., 2007). These interventions typically are delivered to students in small groups and are designed to supplement the activities of Tier 1 intervention for 20-30 minutes per day. The instruction approach is variable, but typically involves implementation of a standard protocol intervention when the goal is for the student to acquire new skills (Fuchs & Fuchs, 2007). Standard protocol interventions are practices of providing academic interventions that are delivered in accordance with a predetermined plan, often in small groups, and supported by rigorous empirical evidence (Glover & DiPerna, 2007). The standard protocol model draws on a body of randomized control group research that has identified specific interventions that have proven to improve the achievement of most students. Students participating in Tier 2 interventions continue to participate fully in the general curriculum and in all of the activities provided at that level of intervention (Vaughn & Roberts, 2007). Tier 2 interventions are typically delivered by supplemental teaching staff such as reading specialists, but are sometimes delivered by the same classroom teacher providing Tier 1 instruction. The assessment system is typically the same as in Tier 1, utilizing curriculum based measures of key learning benchmarks. However, the frequency of progress monitoring is increased from three times per year to 2-3 times monthly to increase the
sensitivity to detect response to the intervention (Brown-Chidsey & Steege, 2005; Burns & Coolong-Chaffin, 2006).

School psychologists can play several important roles in the activities at Tier 2. The role of the school psychologist at Tier 2 is similar to the role at Tier 1, with activities primarily focused on consultation in the administrative process of selecting appropriate empirically based interventions, consultation for valid and reliable assessment, and data-based decision making (Burns & Coolong-Chaffin, 2006; Burns et al., 2007). The school psychologist is a professional with a high level of expertise in identifying appropriate empirically-based standard protocol interventions (Kratochwill & Shernoff, 2004; Kratochwill & Stoiber, 2000; Kratochwill et al., 2000), and particularly those designed for Tier 2 implementation (Fuchs & Fuchs, 2007). School psychologists also offer a high level of expertise in using assessment data to determine students’ response to intervention. School psychologists possess expertise in data-driven decision making, such as classifying response to intervention as responders or nonresponders using a dual discrepancy model (Fuchs & Fuchs, 1998, 2007). Students who have positively responded to the Tier 2 interventions can return to the level of universal intervention with continued progress monitoring. School psychologists are also experienced in using assessment data to identify students needing the most intensive interventions at Tier 3. Use of the dual discrepancy model involves identifying those students whose achievement of learning benchmarks and slope of improvement are both at least one standard deviation below that of peers (Fuchs & Fuchs, 2007).
Tier 3

The third tier, intensive, represents approximately 5% of students who have not adequately responded to Tier 2 targeted interventions, and who therefore require highly individualized and tailored interventions, including students with disabilities (Burns et al., 2007). These interventions typically occur in very small groups of one to three students for fifty minutes or more each day if the interventions are so individually designed that they replace universal interventions (Vaughn & Roberts, 2007). At this level, interventions are most likely to be provided by highly trained specialists such as reading specialists, special education teachers, or behavior interventionists (Fuchs & Fuchs, 2007; Vaughn & Roberts, 2007). In contrast to interventions at Tier 2, there is less reliance on standard protocols for instruction. Implementation of a problem-solving model approach allows for the design of interventions tailored to the individual needs of students (Burns & Ysseldyke, 2005). The problem solving model is a series of steps designed to assist a team of educators in making intervention decisions when attempting to move a student from a current state to a future desired state (Deno, 2002). A summary of the problem-solving model is found in Table 1. At this level, highly individualized interventions are provided and are often based on a scientifically validated methodology that necessitates specific and intensive professional development for teachers (Burns & Coolong-Chaffin, 2006; Sugai & Horner, 2006). Student progress is monitored on a weekly or bi-weekly basis with Tier 3 interventions, utilizing curriculum-based measurement procedures (Reschly & Ysseldyke, 2002).
Table 1

*Data-Based Problem-Solving Model*

<table>
<thead>
<tr>
<th>Problem-solving steps</th>
<th>Assessment procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Problem identification</td>
<td>Observing/recording student performance</td>
</tr>
<tr>
<td>2. Problem definition</td>
<td>Quantifying the perceived discrepancy</td>
</tr>
<tr>
<td>3. Designing intervention plans</td>
<td>Exploring alternative goals and solution hypotheses</td>
</tr>
<tr>
<td>4. Implementing the intervention</td>
<td>Monitoring intervention fidelity and data collection</td>
</tr>
<tr>
<td>5. Problem solution</td>
<td>Requantifying the discrepancy</td>
</tr>
</tbody>
</table>


School psychologists play a critical role in designing, selecting, and implementing, and evaluating interventions at Tier 3. School psychologists typically are leaders of problem-solving teams as they implement the steps of the problem-solving model (Burns & Coolong-Chaffin, 2006; Deno, 2002). School psychologists also have significant expertise in addressing learning difficulties of individual students and designing individualized interventions. They are critical to the problem-solving process as they provide leadership in designing and interpreting assessment data systems (Daly, Chafouleas, & Skinner, 2005; Shapiro, 2004). Much assessment at Tier 3 involves single-subject research paradigms and brief experimental analysis, activities for which school psychologists are uniquely qualified. These methodologies are based on a hypothesis testing approach in which individual student responses to intervention are
systematically compared to their baseline performance, and often withdrawing the interventions to return to baseline condition (Berg, Wacker, & Steege, 1995; Brown-Chidsey & Steege, 2005).

*Description of Program Evaluation Processes at Each Tier*

Program evaluation at each tier of this model typically involves different research designs and evaluation methods. Evaluation of interventions used at Tier 1 generally does not present a challenge of research methodology (Burns, 2007). At Tier 1, large scale experimental designs and descriptive research methods are appropriate because any student can be initially assigned to the intervention condition or to an alternative control group that receives traditional instruction (Shadish, Cook, & Campbell, 2002). Random assignment to treatment or control groups can permit the researcher to determine effectiveness of the core curriculum program in comparison to other traditional programs of instruction (McCaffrey, Ridgeway & Morral, 2004). Program evaluation can be conducted using classic large-group statistical methods utilizing either curriculum-based measures or standardized state performance measures (Burns, 2007; VanDerHeyden & Burns, 2005). Utilization of these classic large-group, randomized control studies is considered the “gold standard” in program evaluation (U.S. Department of Education Institute of Education Sciences, 2003).

Program evaluation of Tier 2 interventions typically presents a research challenge for practitioners and researchers and is the focus of the current study. Because students are assigned to these interventions as a result of poor response to quality instruction at the universal level, the sample of students participating in Tier 2 interventions is inherently
biased (Burns, 2007; Burns et al., 2007; Newman et al., 2006). Individualized interventions are highly time-consuming if they are implemented and evaluated effectively. If they are utilized with up to 20% of the student population, the problem-solving team members will not have the resources to work effectively with that size population (Burns, 2007). As a result, interventions and their evaluation at Tier 2 must be carried out without reliance on an individual problem-solving approach. Instead, students should be assigned to standard protocol interventions that are delivered in small groups (Fuchs & Fuchs, 2007; Glover & DiPerna, 2007). School psychologists working to develop system-wide interventions at Tier 2 can play an important role in evaluating the effectiveness of those interventions on the target group (Burns & Coolong-Chaffin, 2006).

A body of research exists identifying the inherent sample biases among students assigned to Tier 2 interventions in applied settings (McMaster, Fuchs, Fuchs, & Compton, 2005). For example, students participating in Tier 2 interventions such as Title 1 reading programs are assigned specifically because they have low reading test scores. Assigning students to a treatment group based upon reading achievement test scores would create a biased treatment group of poor readers (U.S. Department of Education Institute of Education Sciences, 2003). Research has also shown that racial minorities, economically disadvantaged students, and English language learners have historically been over represented in reading intervention programs such as Title 1 (Carlisle et al., 2006). These factors also represent sources of inherent bias in the samples under study.
The presence of a known systematic bias presents a critical challenge to school psychologists acting in the role of intervention program evaluators. The most common methods of controlling for this bias are the use of systematic regression techniques or Analysis of Covariance (ANCOVA). The ANCOVA technique allows the researcher to compare the gains of groups by adjusting the means to control for the selection bias. For example, ANCOVA was utilized by Gettinger and Stoiber (2007) in their program evaluation of the early literacy initiative Exemplary Model of Early Reading Growth and Excellence (EMERGE). Adjusted achievement scores were compared after controlling for the entering achievement differences between the intervention group and the control group when group membership eligibility was significantly correlated with entering achievement. Adjusted means approaches such as regression analysis and ANCOVA have been used in similar ways in applied program evaluations of reading interventions focusing on a variety of independent variables such as levels of teacher training (Modaressi & Wolanin, 2007), improving reading motivation through instruction in improving relevant context (Aarnoutse & Schellings, 2003), supplementing intensive literacy intervention programs with the Rainbow repeated reading program (Wheldall, 2000), varying levels of instruction in heuristics among high school students (Nokes, Dole, & Hacker, 2007), increasing levels of intervention duration and intensity among poor readers in elementary school (Vaughn et al., 2009), and the effect of the Talent Development high school program on reading skills among high poverty students (Balfanz, Legters, & Jordan, 2004).
Evaluating the progress of students receiving the most intensive individualized interventions at Tier 3 is most appropriately conducted using single subject research designs as part of the problem-solving process (Burns, 2007; Burns & Coolong-Chaffin, 2006; Burns et al., 2007; Sugai & Horner, 2006; Walcott & Riley-Tillman, 2007). Although the effectiveness of an intervention may have been validated by prior EBI research as being effective with 90% of children studied, in practice the school psychologist may be working with a child who falls in the remaining 10% who are not responders. It is the responsibility of the school psychologist to determine if the intervention is effective for a particular child. By applying single-subject research designs, the school psychologist can make these highly individualized determinations (Walcott & Riley-Tillman, 2007).

Single-subject research designs utilize a hypothesis testing procedure to test intervention effectiveness on an individual (Kazdin, 1982). In contrast to between-group designs that compare the performance of an experimental and control group, single-subject designs compare the functioning of a single subject through a series of experimental and control phases (Brown-Chidsey & Steege, 2006; Kazdin, 1982; Polaha & Allen, 1999; Steege, Brown-Chidsey, & Mace, 2002). In the baseline phase (A), performance data are collected on the target behavior (e.g. reading fluency, or words per minute) for several days before interventions are introduced, providing the school psychologist with a basis of comparison similar to a control condition. In the intervention phase (B), the intervention is implemented and performance data are collected daily using the same measurement system that was used in the baseline
condition (i.e. words per minute). The simplest single-subject research design, case study, involves only two phases of baseline and intervention (A-B) (Kazdin, 1982). In a reversal design, successive phases are added, alternating baseline and intervention conditions (A-B-A-B). A multiple baseline design may be used to determine the intervention’s effectiveness on a target behavior and a related generalized behavior, such as spelling test words and spelling in journal writing tasks (Brown-Chidsey & Steege, 2006). In such a design, baselines for more than one behavior are taken in the initial phase and followed through the course of one or more interventions. Another common variation of a single-subject design is a changing conditions design, in which successive phases in which interventions are expanded (A-B-B+C).

After data are collected through the phases of the chosen research design, the school psychologist must interpret the data to determine the student’s response to intervention. Behavioral data through all phases of baseline and intervention are first recorded in a linear graphic form (Steege et al., 2002). The data may be analyzed by a process of visual inspection or statistical analysis (Alberto & Troutman, 2003). In visual inspection, the school psychologist considers the observed changes in level, slope, and variability in baseline and treatment data (Brown-Chidsey & Steege, 2006). In statistical analysis, the school psychologist considers the changes in the mean of collected data points through successive treatment and baseline phases (Alberto & Troutman).

Indecision regarding the selection of an appropriate intervention can paralyze a problem-solving team working at Tier 3. Without information generated from single subject experimental design, problem solving teams often select an intervention for
nonspecific reasons, i.e., because it has recently been validated in a group study, because it is familiar to a member of the team, because anecdotal evidence (“It worked with another student”), or because of mere intuition (“It looks like it should work”)(Brown-Chidsey & Steege, 2006). School psychologists play a critical role in applying these single-subject experimental methodologies in assisting problem solving teams to select interventions that have been individually validated.

Literature Review Related to School Psychologist’s Role in Tier 2 Program

Evaluation

Historically, the primary role of school psychologists has been individual child study with the objective of developing individual interventions and special education eligibility decisions. (Burns et al., 2006; Burns, 2007; Fagan & Wise, 2007; Reschly & Ysseldyke, 2002; Ysseldyke et al., 2006). In the next decade, school psychologists are envisioned playing an increasing role in addressing the academic achievement of students with and without disabilities in the general education environment. There are many forces from within and from outside the field of school psychology contributing to a role shift toward addressing academic achievement among students in the general education population. An example of a force outside the field of school psychology centers on the mandates stemming from NCLB. NCLB has made schools accountable for the quantifiable achievement of all students, especially those who have historically “slipped through the cracks” of the educational system. In order to address the needs of a diverse population, schools will need hands-on consultants to address these requirements for student improvement, consultants who have expertise in the instruction process, the
learning process, assessment, problem solving, statistics and research methodology. School psychologists are uniquely trained in this skill set, making them indispensable contributors in this school improvement effort (Fagan & Wise, 2007).

The National Association of School Psychologists (NASP) has produced a guiding document for university-based school psychology training programs that addresses the changing roles in the profession, the *Blueprint for Training and Practice III* (Ysseldyke et al., 2006). This document identifies the primary outcomes school psychologists should be striving toward, the delivery system within which school psychologists should function, and the specific domains of competence which are within the role and function of school psychologists as they engage in practice at multiple levels of assessment and intervention. *Blueprint III* suggests that school psychologists should be working towards two primary outcome goals: (a) improving competencies for all students, and (b) building and maintaining the capacities of systems to meet the needs of all students. School psychologists should work within a three-tiered delivery system that serves the needs of all learners. Specific foundational and functional competencies promoted by *Blueprint III* relevant to the task of closing the achievement gap include enhancing the development of cognitive and academic skills in all students, systems-based service delivery, and data-based decision making and accountability. Table 2 provides a summary of foundational and functional competencies identified in *Blueprint III* that relate to this expanded role of promoting higher achievement for all students.
<table>
<thead>
<tr>
<th>Domain</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhancing the Development of Cognitive and Academic Skills</td>
<td>School psychologists help schools develop challenging but achievable cognitive and academic goals for all students, taking into account the need to adjust expectations for individual students, or to implement alternative ways to monitor or assess individual student progress toward goal or standards accomplishment.</td>
</tr>
<tr>
<td>Systems-Based Service Delivery</td>
<td>School psychologists should provide leadership in developing schools as safe, caring, and inviting places in which there is a sense of community, in which contributions of all persons are valued, in which there are high expectations of excellence for all students, and where home-school-agency partnerships are valued.</td>
</tr>
<tr>
<td>Data-Based Decision Making &amp; Accountability</td>
<td>School psychologists should be good problem solvers who collect information that aids in understanding problems, making decisions about appropriate interventions, assessing educational outcomes, and making accountability decisions.</td>
</tr>
</tbody>
</table>

School Psychologists’ Role in Program Evaluation Suggested by Blueprint Outcome

Goals

Schools exist for the purpose of preparing all students to become more competent in two broad categories of skill; (a) academic and cognitive competencies such as reading and math skills, and (b) social-emotional competencies such as getting along with others and the ability to cope with frustrations (Ysseldyke et al., 2006). School psychologists are educational professionals whose primary goal is to seek the development of these outcomes for all students. The training competencies described in Blueprint III (Ysseldyke et al., 2006) suggest that, to help schools achieve these outcome goals, school psychologists should be utilized as instructional consultants who assist teachers and parents in identifying effective instructional practices in the school setting, and determining how students are learning best. School psychologists are also mental health practitioners who can guide teachers and parents to foster safe and supportive home and school environments. This outcome goal emphasizes the school psychologist’s role in helping determine what effective practices are for all children, not just students with disabilities.

This outcome goal for school psychologists suggests a major shift in the role school psychologists play in schools. Prior role descriptions for school psychologists focused on individual child study and remediation of deficits in individual children, whereas contemporary role definitions suggest that school psychologists are focused on developing and assessing the effectiveness of programs for large groups of students (Bradley-Johnson & Dean, 2000; Fagan & Wise, 2007; Reschly & Ysseldyke, 2002;
Interventions that have an impact on larger groups of students without disabilities typically have a standardized program component which allows for efficient and consistent interventions delivered to students in small groups, and these standardized small group interventions typically exist in Tier 2. School psychologists should play important roles in selecting and evaluating the effectiveness of these types of interventions (Burns & Cooling-Chaffin, 2006).

The second major outcome goal for school psychologists described in Blueprint III is to help build and maintain the capacities of educational systems (Ysseldyke et al., 2006). In order for schools to achieve the goal of improving academic and behavioral outcomes, schools must recognize that students learn and grow within a larger learning system. Components of the system must work together efficiently with focused action to maximize the capacity of the system to reach all students. This means that school psychologists should play an important role in consulting with school leaders in many activities that promote continuous improvement of systems. The role of systems consultant is described in Blueprint III to include the ongoing evaluation of the functioning of the educational system for all learners. In order for school psychologists to influence the capacity of the system to grow, it is necessary for them to share in the vision of broad outcome targets, act as consultants in developing strategies to reach those targets, and provide expertise in the ongoing evaluation of the effectiveness of those strategies. Evaluation of learning outcomes for all students indicates that school psychologists will play an increasing role in conducting program evaluation activities such as designing internal research evaluation protocols, analyzing performance data, and
interpreting these data in light of the district’s achievement targets. The end result of these activities is for the school psychologist to help the system use its personnel, fiscal resources, and activities efficiently in pursuing improved outcomes for all students (Ysseldyke et al., 2006).

_School Psychologists’ Role in Program Evaluation Suggested in Blueprint Domains of Competence_

Accomplishing the two outcome goals (i.e., improving competencies for all children and building capacity of educational systems) suggests a new vision for the field of school psychology. This vision of an emerging role for school psychologists places a premium value on not only designing but also conducting evaluation of both individual intervention programs and standardized programs designed for implementation among groups of students with and without disabilities. Three specific functional competencies promoted by _Blueprint III_ relevant to the task of evaluating programs aimed at closing the achievement gap include (a) enhancing the development of cognitive and academic skills in all students, (b) systems-based service delivery, and (c) data-based decision making and accountability (National Association of School Psychologists, 2009; Ysseldyke et al., 2006). The role envisioned for today’s school psychologist in Blueprint III demands that practitioners in the field move far beyond knowledge about research methodology skills to applying these skills in conducting program evaluations of the systems of intervention put into place for both individual students and groups of students.

The functional competency of data-based decision making and accountability is presented in _Blueprint III_ as a process that permeates all activities of school
psychologists. Practitioners in the profession are intended to be continuous collectors of information that is relevant to identifying effective interventions and evaluating educational outcomes. They should use this information systematically on behalf of all children, not just individual students with disabilities as school psychologists have historically been viewed (Fagan & Wise, 2007; Reschly & Ysseldyke, 2002). School psychologists have been trained extensively to be data-based decision makers in the activities required for individual child study. They have also received extensive training in statistics, research methods, and program evaluation, and it is this training that enables them to be data-based decision makers on behalf of school systems, programs, and classroom environments (National Association of School Psychologists, 2008).

The functional competency of systems-based service delivery suggests that school psychologists operate in the role of helping all students become more competent by affecting the environments and systems in which students receive their education. They should have the competency to design programs and systems in ways that promote learning and prevent problems. Using a 3-tiered model of academic intervention, the school psychologist has the competence to develop programs, instructional interventions, and problem-solving teams for determining the levels of intervention students require (Fagan & Wise, 2007; Silva, 2003; Ysseldyke et al., 2006).

Implementation of interventions across a multiple tier model suggests that school psychologists involved in the systemic process could have a variety of roles at each tier of intervention (Tilly, 2002). Burns and Coolong-Chaffin (2006) suggest that school psychologists working to enhance Tier 2 interventions would be involved in identifying
reliable and valid assessment processes, data-based decision making practices, interpretation of student achievement data, monitoring of implementation integrity, and enhancing home/school collaboration. In particular, one of the most obvious roles for school psychologists is in consulting with administrators to design the assessment system used to assign students to tiers of intervention (Burns, Dean & Klar, 2004; Fuchs, 2003; Gresham, 2002). The assessment system must be reliable, valid, appropriately sensitive to gradual changes in student achievement, and in a form useful for conducting program evaluation.

**Literature Review Related to Evidence-Based Tier 2 Reading Interventions**

Although there is wide agreement that students should be assigned to Tier 2 interventions based upon limited success at the universal level of intervention (Burns et al., 2006; Vaughn & Roberts, 2007) and there is wide agreement that the percentage of the student population that should receive Tier 2 interventions should be roughly 15-20% (Burns et al., 2007; Griffiths et al., 2007; Sugai & Horner, 2006), there is less agreement about what exactly constitutes a Tier 2 intervention (Burns et al., 2006). This section will include a review of literature related to successful Tier 2 interventions. First, research on the effectiveness of both commercially prepackaged and academically prepared reading interventions will be reviewed. Service delivery factors that may affect the success of Tier 2 interventions will be discussed for both commercially prepackaged and academically prepared interventions. Lastly, the features common to successful Tier 2 reading interventions will be described and summarized.
Selection of Evidence-Based Tier 2 Interventions

Selection of evidence-based Tier 2 interventions for discussion in this section was made using two approaches. First, the reviews available through the What Works Clearinghouse (U.S. Department of Education Institute of Education Sciences, 2009) website’s beginning reading section was accessed, and interventions were selected in the outcome areas of general reading achievement, reading comprehension, alphabetes, and reading fluency. The interventions presented in this review are those that were reported with positive effects or potentially positive effects, and have been designed for implementation in 15-20% of the population or in small groups (i.e., matching key features of Tier 2 interventions). The review process and application of the review protocol at What Works Clearinghouse produces a set of ratings for each intervention in the following domains: (a) improvement index, (b) evidence rating, and (c) extent of evidence. The improvement index represents the difference between the mean percentile ranks of students in the intervention group compared to students in the comparison group in terms of their relative levels of performance on the outcome variable. Interventions reviewed by the Clearinghouse are ranked overall based on the improvement index on values ranging from -50 to +50. The evidence rating considers the factors of quality of research design, statistical significance of between-group differences, the magnitude of between-group differences, and the consistency of replication of results across studies. A minimum effect size of .25 based on Hedges g computation of pooled variance (Hedges, 1981) is the threshold for a “potentially positive” designation. This minimum effect size typically translates to a 10 percentile rank increase in the improvement index (U.S.)
Department of Education Office of Information Science, 2009). The extent of evidence refers to the number of studies and the number of students who participated in those studies. Interventions are classified as having either small or medium to large extent of evidence. A list of targeted interventions that have met the standards of rigorous evidence at What Works Clearinghouse is presented in Table 3. Only those interventions that have been validated in studies incorporating a randomized control group design meet this standard of evidence.

Second, interventions were selected from reviews available in the peer-reviewed literature that utilized selection criteria developed within the profession of school psychology for evidence-based interventions (Kratochwill & Shernoff, 2004; Kratochwill & Stoiber, 2000; Kratochwill et al., 2000; Shernoff et al., 2003). These selection criteria are described in the review manual of the Task Force on Evidence-Based Interventions in School Psychology established by APA Division 16 (Task Force on Evidence-Based Intervention in School Psychology, 2003).
<table>
<thead>
<tr>
<th>Intervention</th>
<th>Outcome</th>
<th>Improvement</th>
<th>Evidence</th>
<th>Extent of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Recovery</td>
<td>General Reading</td>
<td>32</td>
<td>Positive Effects</td>
<td>Med/Large</td>
</tr>
<tr>
<td></td>
<td>Fluency</td>
<td>46</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td>Comprehension</td>
<td>14</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td>Alphabetics</td>
<td>36</td>
<td>Positive Effects</td>
<td>Med/Large</td>
</tr>
<tr>
<td>SMART</td>
<td>Fluency</td>
<td>17</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td>Comprehension</td>
<td>14</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td>Alphabetics</td>
<td>16</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td>Earobics</td>
<td>Fluency</td>
<td>15</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td>Alphabetics</td>
<td>25</td>
<td>Positive Effects</td>
<td>Small</td>
</tr>
<tr>
<td>PALS</td>
<td>Fluency</td>
<td>13</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td>Comprehension</td>
<td>13</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td>Alphabetics</td>
<td>19</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td>EIR</td>
<td>Comprehension</td>
<td>18</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td>Alphabetics</td>
<td>36</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td>Failure-Free Reading</td>
<td>Comprehension</td>
<td>10</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td>DaisyQuest</td>
<td>Alphabetics</td>
<td>23</td>
<td>Positive Effects</td>
<td>Small</td>
</tr>
</tbody>
</table>
Table 3 (Continued)

*Targeted Interventions Meeting What Works Clearinghouse Rigorous Evidence Criteria*

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Improvement Outcome</th>
<th>Improvement Index</th>
<th>Evidence Rating</th>
<th>Extent of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiPS</td>
<td>Alphabetics</td>
<td>17</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td>Read, Write &amp; Type</td>
<td>Alphabetics</td>
<td>8</td>
<td>Potentially Positive</td>
<td>Small</td>
</tr>
<tr>
<td>FastForWord</td>
<td>Alphabetics</td>
<td>8</td>
<td>Positive Effects</td>
<td>Small</td>
</tr>
</tbody>
</table>

Improvement Index = mean percentile gain for intervention group compared to controls

Evidence Rating = considers research design, level of statistical significance, mean gain, and consistency across studies

Extent of Evidence = considers number and sizes of studies

Many commercially prepackaged Tier 2 reading interventions are available on the market. This section reviews only those prepackaged interventions for which peer-reviewed validation research incorporating a randomized control group is available. One such program is Peer-assisted Learning Strategies (PALS; Fuchs, Fuchs, & Burish, 2000), which was designed to incorporate a system of classwide peer tutoring to supplement classroom reading instruction in kindergarten through third grade. All students continue in the universal reading curriculum and receive supplemental reading intervention through PALS. In PALS, students are each assigned a reading partner and engage in paired reading activities. Students alternate in roles of Player and Coach as they orally read portions of text selected by the teacher. The student in the Coach role listens to the Player read orally and provides corrective feedback on the accuracy of the Player’s reading. Students engage in the reading intervention for 30 minute sessions 3 times per week. Validation research on PALS has demonstrated positive effect sizes ranging from .40 to .49 (McMaster, Fuchs, Fuchs, & Compton, 2005). Effect sizes of at least .20 are considered small, effect sizes of at least .50 are considered medium, and effect sizes of at least .80 are considered large (Hedges, 1981; Cohen, 1977). What Works Clearinghouse reports potentially positive effects for PALS in the outcome areas of fluency, comprehension, and alphabetics skills.

The Reading Recovery program was developed in New Zealand (Clay, 1993) and has been widely available in the United States through its association with the Ohio State University since 1991. Reading Recovery is an intervention of one-to-one tutoring
targeted at the lowest performing 20% of readers. All students continue in the universal reading curriculum, while the targeted students receive supplemental reading intervention through Reading Recovery. Students participate in individual tutoring sessions daily for 30 minutes over a 12-20 week period. Intervention is delivered by licensed teachers who have been specifically trained in the Reading Recovery protocol that incorporate components of phonemic awareness, phonics, vocabulary, fluency, comprehension, writing, motivation, oral language and independence. Each interactive tutoring session consists of reading both familiar and novel stories, manipulating letters and words, and writing and assembling stories. *What Works Clearinghouse* reports positive effects for Reading Recovery based upon peer reviewed literature in the outcome areas of general reading development and alphabetics (Baenen, Bernhole, Dulaney, & Banks, 1997; Pinnell, DeFord, & Lyons, 1988). Potentially positive effects are reported in the outcome areas of fluency and comprehension. A recent meta analysis of available research on Reading Recovery reported positive gains for students who had participated in the full 20 week intervention (D’Agostino & Murphy, 2004).

Start Making a Reader Today (SMART; Oregon Children’s Foundation, 1992) is an intervention program widely used in Oregon with students in kindergarten through second grade. The program utilizes volunteer tutors to provide supplemental reading instruction to students at risk of reading failure. The intervention is targeted in high poverty schools. Adult tutors who have completed a standard two hour training program provide supplemental instruction in a one-to-one ratio during two 30 minute sessions per week for a full school year. Tutors assist individual students in four activities: reading to
the student, reading simultaneously with the student, re-reading with the student, and then asking probing comprehension questions. All students continue in the universal reading curriculum, while the targeted students receive supplemental reading intervention through SMART. In addition, each student participating in SMART is given two new books each month to encourage family reading activities. *What Works Clearinghouse* reports potentially positive effects for SMART based upon peer reviewed literature in the outcome areas of alphabetics, fluency, and comprehension (Baker, Gersten, & Keating, 2000).

**Earobics** is an interactive software package designed to provide individualized instruction in literacy skills to students in pre-kindergarten through third grade (Houghton Mifflin Harcourt Learning Technology, 1995). Because is designed as a supplemental instruction program, all students continue in the universal instruction of the general curriculum. Targeted students utilize the Earobics software to supplement the development of early literacy skills. Students interact with animated characters through a visual and musical software interface that automatically adjusts the level instruction to the responses of the learner. In each level of instruction, students are exposed to lessons in sound recognition and sound blending, rhyming, and discriminating phonemes within words. Supplemental materials in hard copy, such as picture/word cards, letter-sound decks, big books, little books and leveled readers permit the teacher to provide reinforcement through human interaction. *What Works Clearinghouse* reports potentially positive effects for Earobics based upon peer reviewed literature in the outcome area of
reading fluency, and positive effects in the outcome area of alphabetics (Gale, 2006; Cognitive Concepts, Inc., 2003).

Early Intervention in Reading (EIR; Taylor, Frye, Short, & Shearer, 1991) is a Tier 2 intervention program designed to provide supplemental instruction to students in kindergarten through fourth grade who are at risk of failing to learn to read. In grades kindergarten through second, the program contains both universal instruction and supplemental instruction components, with the supplemental components designed for use with approximately 20% of students who are not making satisfactory progress in whole-class instruction. In grades three and four, the program is designed for supplemental instruction only. The program is designed to be implemented by classroom teachers with the expectation that they will be able to provide targeted supplemental instruction to small groups of 5-7 students, four days per week for 20 minutes. The program utilizes picture books to emphasize reinforcement of phonics, phonemic awareness, and contextual analysis. Repeated reading and repeated writing strategies are employed to nurture fluency skills. Teachers utilizing the intervention receive intensive professional development over a nine month period in which they attend workshops and engage in internet-based instruction. What Works Clearinghouse reports potentially positive effects for EIR based upon peer reviewed literature in the outcome areas of alphabetics and reading comprehension (Taylor, 2001; Taylor, Frye, Short, & Shearer, 1991).

Failure Free Reading (Lockavich, 2008) is a Tier 2 intervention program designed to provide supplemental instruction to the approximate 15% of students in kindergarten
through twelfth grade who have not responded successfully to conventional reading instruction. It is promoted as a language development program that focuses on improving fluency, vocabulary, word recognition and comprehension. Instruction is delivered to small groups or individually, and emphasizes repeated exposure to reading passages, use of predictable sentence structures, and stories that minimize the need for prior knowledge. Students engage in repeated reading of passages, repeated hearing of passages, and exposure to repeated patterns in language. The program is designed upon the premise that repetition is engaging, not boring, for struggling readers. Failure Free Reading utilizes a standard protocol approach that combines systematic scripted teacher instruction, talking computer software, independent reading assignments and workbook exercises. *What Works Clearinghouse* reports potentially positive effects for Failure Free Reading based upon peer reviewed literature in the outcome area of reading comprehension (Torgesen et al., 2006).

DaisyQuest (Torgesen, 1997) is a stand-alone computer software program that can be used in the classroom as a supplement to the universal curriculum for children ages three to seven. The software engages students in a story line that involves searching for a dinosaur named Daisy and her lost eggs. The development of phonological awareness is targeted through multiple activities that isolate the components of word rhyming. Tutorial instruction is provided in each skill, followed by sets of practice items. Students practice the following tasks: (a) recognizing words with the same beginning, middle, and ending sounds, (b) building words from a set of phonemes, and (c) counting the sounds in words. As students master the tasks at each level, they are rewarded with
clues to the location of Daisy and her eggs. An adaptive assessment tool is included with the software that provides statistical feedback for teachers and parents. Students in the peer-reviewed studies used the DaisyQuest software for 15 to 32 sessions of 20 to 25 minutes. *What Works Clearinghouse* reports positive effects for DaisyQuest based upon peer reviewed literature in the outcome area of alphabetics skills (Barker & Torgesen, 1995; Foster, Erickson, Foster, Brinkman, & Torgesen, 1994; Mitchell, & Fox, 2001).

The Lindamood Phonemic Sequencing (LiPS) program (Lindamood & Lindamood, 1998) is designed to teach students decoding skills through identifying individual sounds and blends in words. The foundational activity is engaging the students in discovering the motor actions of the lips, tongue, and mouth used to produce each sound in the English language. After students acquire the skills of producing, labeling, and organizing all of the sounds, they apply the oral aspects of each sound as they sequence them in reading and spelling activities. The program is designed for students in kindergarten through third grade or for struggling readers of any age. Implementation of the intervention is through small group or individual instruction, one hour daily for four to six months. Although the program is designed as a targeted intervention to supplement the universal reading intervention, the recommended time investment suggests that the program may also be suited as a Tier 3 intervention. *What Works Clearinghouse* reports potentially positive effects for LiPS based upon peer reviewed literature in the outcome area of alphabetics skills (Torgesen, Wagner, Rashotte, & Herron, 2003).
The Read, Write & Type! Learning System (Read, Write & Type!; Herron & Grimm, 1994) is a software program that emphasizes learning to read through the writing process. The program is targeted at students aged six to nine years or who are older struggling readers, and is designed as a supplement to the universal curriculum reading intervention. The main goal of the program is to instruct students in the awareness of the 40 different phonemes in the English language and to associate each phoneme with a single letter or combination of letter keys on the keyboard. Using a game-like set of activities with animated characters, the software directs students through an instructional sequence that begins with students saying and typing individual sounds and their letter representations. Later, students combine their phoneme keystrokes to construct words, sentences, and stories. Students utilize the software independently and do not require explicit guidance from a trained adult. What Works Clearinghouse reports potentially positive effects for Read, Write & Type! based upon peer reviewed literature in the outcome area of alphabetics skills (Torgesen, Wagner, Rashotte, & Herron, 2003).

Fast ForWord (Tallal, 2000) is a computer software system designed to strengthen the cognitive prerequisite skills necessary for fluent reading and other learning tasks. The program targets the cognitive domains of memory, attention, processing and sequencing through the use of language-based activities. Students listen through headphones and respond to probes to improve their listening accuracy, phonological awareness and understanding of language structures. Later, these skills are applied to the process of sound-letter comprehension and word recognition. Students utilize the software for 30 to 100 minutes 5 days per week for four to 16 weeks, making this intervention suitable for
either a Tier 2 or Tier 3 implementation depending upon intensity of application. *What Works Clearinghouse* reports positive effects for Fast ForWord based upon peer reviewed literature in the outcome area of alphabetics skills (Borman & Benson, 2006; Scientific Learning Corporation, 2006; 2005a; 2005b; 2005c).

**Additional Evidence-Based Interventions in School Psychology Literature**

Many university studies, funded by federal grants in cooperation with state education departments, have developed intervention programs that incorporate principles promoted by the National Reading Panel. These principles include direct instruction in phonemic awareness, phonics, fluency, vocabulary, and text comprehension (National Reading Panel, 2000). Many of these studies meet the standards of evidence described in the review manual of the Task Force on Evidence-Based Interventions in School Psychology established by APA Division 16 (Task Force on Evidence-Based Intervention in School Psychology, 2003). In a recent review of Tier 2 reading interventions, 13 programs met selection criteria set by Vaughn and Wanzek (2007) that included interventions targeted in grades K-3, involved 100 sessions or more, targeted early literacy, were school-based, and included students at-risk of learning problems. They reported effect sizes of -.43 to 1.33 among various levels of these programs, indicating that some were highly effective while others seemed to be counterproductive. The selection criteria for this review differed from *What Works Clearinghouse* criteria in research design standards. Most of the studies selected incorporated a treatment and comparison group design that was not randomly assigned.
One of the most promising studies reviewed by Vaughn and Wanzek was the Sound Partners and Thinking Partners program (Vadasy, Sanders, Peyton, & Jenkins, 2002) in which one-to-one tutors utilize phonics-based instruction to assist first through third grade English language learners and students at risk for reading failure. The program is designed for implementation by trained paraprofessionals in the school setting with small groups of students for 30 minutes, four times per week. The primary feature of instruction in Sound Partners is explicit teaching of phonological skills applied to decodable text. Effect sizes reported for Sound Partners were .73 for phonemic decoding, .88 on a spelling measure, and .92 on a word reading measure. A follow-up study (Vadasy, Sanders, & Peyton, 2005) of Sound Partners reported effect sizes of 1.33 for word attack skills and 1.12 for word reading. Both studies employed quasi-experimental designs and used an adjusted means (ANCOVA) technique to adjust for differences in entering reading skills of the control group.

Torgesen et al. (1999) also utilized a one-to-one tutoring intervention that emphasized explicit phonics instruction with students in grades K-2 called Phonological Awareness Plus Synthetic Phonics (PASP). This intervention is designed for implementation by highly trained teachers in an individual instruction setting. Torgesen et al. provided direct instruction in phonological skills and instruction in oral motor awareness to students with reading delays. They reported an effect size of 1.21 on a measure of sight word recognition, 1.04 for word attack, and .93 for nonword reading in this randomized control study.
Jenkins, Peyton, Sanders and Vadasy (2004) investigated the effects of using more phonetically decodable text and less phonetically decodable text in a phonics tutoring intervention. They utilized a one-to-one tutoring ratio among at-risk readers in first grade. Paraprofessional educators received training as individual tutors to provide explicit phonics instruction. Following an intervention protocol requiring 30 minute sessions 4 times per week for 25 weeks, an effect size of .81 was reported on the oral reading accuracy measure for more decodable text, while an effect size of .51 was reported for less decodable text in relation to the control group. This study employed an adjusted means (ANCOVA) technique to adjust for differences in entering reading skills of the control group.

The Early Steps program (Morris, Tyner, & Perney, 2000) is an intervention that has many similarities to Reading Recovery. Highly trained tutors provide explicit instruction in one-to-one instructional settings, 30 minutes per day for at least 20 weeks, to students who have not responded favorably to the universal curriculum. Early Steps differs from Reading Recovery in its emphasis on orthographic instruction in isolation before the application to reading text. Effect sizes of .76 were reported for word attack; .83 for developmental spelling; and .79 for passage reading. This study used a matched control group design, but did not utilize an adjusted means approach to adjust for differences in the baseline skills of the control group.

Help One Student to Succeed (HOSTS; 1999) is a Tier 2 reading intervention that incorporates a structured mentoring program by matching students from grades K-12 to one-to-one adult mentors. Diagnostic assessments and progress monitoring are key
components of the program as is the incorporation of proprietary software. The HOSTS program is designed to supplement the core curriculum used in the school district. Adult tutors are identified in the community and must be trained to administer the mentoring instruction. Validation research on HOSTS has demonstrated positive effect sizes of .44 for reading comprehension and .99 for reading fluency (Burns, Senesac, & Symington, 2004). This study employed an adjusted means (ANCOVA) technique to adjust for differences in entering reading skills of the control group.

The interventions described in this section are not an exhaustive listing, but are the best examples of Tier 2 interventions based upon the evidence standards of the U.S. Department of Education and the profession of school psychology. Implementation of an appropriate intervention in an applied setting will require both the selection of an evidence-based intervention and consideration of several local variables, conditions and circumstances that may have an impact on the success of the intervention in the selected setting.

Critical Service Delivery Features

The selection of appropriate, evidence-based reading interventions for use at Tier 2 should also consider several practical components that relate to the manner and quality of service delivery. This section will present a brief summary of treatment fidelity, instructor-to-student ratio, instructional term, and professional status of personnel delivering interventions.
**Treatment Fidelity**

Ultimately, interventions are conducted to improve the academic and behavioral skills of children, not to produce research data. The processes by which interventions are implemented must be flexible enough to respond to differences in students, the schools they attend, and the teachers they employ (Burns et al., 2006). Nevertheless, researchers report that the most significant barrier to identifying effective Tier 2 interventions is inconsistency in the implementation of interventions, called treatment fidelity (Gresham, Gansle, Noell, Cohen, & Rosemblum, 1993). Treatment fidelity is the degree to which a planned intervention is implemented as designed (Telzrow & Beebe, 2002). Of concern are the implementation of classroom interventions as designed, the process by which they are designed (Burns & Ysseldyke, 2005), and the acceptability of those interventions to teachers or other professionals who serve as the direct interventionists (Eckert & Hintze, 2000). Witt and Elliott (1985) conceptualized overall treatment acceptability as comprised of the four sequential and reciprocal elements of treatment acceptability, treatment integrity, treatment use, and treatment effectiveness. School psychologists acting in a consulting role to select, design, and implement Tier 2 interventions must recognize the importance assessing each of these interrelated elements throughout the intervention process.

In order to know the degree to which an intervention is implemented as designed, it is necessary to have tools for measuring treatment fidelity. Methods that have been reported in the literature include participant reports, evaluation of permanent products, and collecting reports from outside sources. Participant reports are reports of
intervention integrity collected and reported by intervention implementers themselves, usually teachers. These self-report data, though easy to collect, are highly unreliable (Wickstrom, Jones, LaFleur, & Witt, 1998). Permanent product records, such as activity sheets, service logs, and intervention scripts can be valuable tools for determining if an intervention is implemented as designed (Telzrow & Beebe, 2002). The most valuable data for integrity checks are reports collected from outside sources. The preferred method for these integrity checks are incorporating direct observations of intervention events by trained observers over multiple trials and settings (Moncher & Prinz, 1991).

School psychologists interested in improving the outcomes for students targeted for Tier 2 intervention should be cognizant of the factors impacting treatment acceptability and fidelity. Factors that have been found to influence overall acceptability are the severity of the referral problem, the selection of evidence-based interventions, selection of interventions that are positive rather than reductive, ease of pragmatic implementation (Cowan & Sheridan, 2003; Eckert & Hinze, 2000). Factors that have been found to influence treatment integrity are use of a standard protocol or step-by-step script for implementation, providing ongoing guided practice and feedback from an experienced interventionist, and incorporating treatment integrity checks from an outside observer (Telzrow & Beebe, 2002).

*Instructor-to-Student Ratio*

Of the peer-reviewed Tier 2 interventions identified by Vaughn and Wanzek (2007), the interventions with the strongest effect sizes utilized the smallest group sizes. The three most effective intervention programs utilized a one-to-one student to tutor
ratio. Based on meta-analyses of intervention research, other researchers have argued that small group instruction models can be equally effective if not more effective than one-to-one models (D’Agostino & Murphy, 2004; Elbaum, Vaughn, Tejero & Watson, 2000). Because financial limitations necessarily limit the number of students that can benefit from interventions, small groups of 4 to 6 students appears to be a ratio that balances intervention effectiveness with intervention efficiency (Burns et al, 2006).

*Instructional Term*

There is great variation in the term and duration of intervention activities presented in the literature. The Reading Recovery standardized intervention for first grade requires 30 minute sessions daily for 12 to 20 weeks (D’Agostino & Murphy, 2004). The Sound Partners standardized intervention that resulted in very significant effect sizes for both word attack and comprehension skills required 30 minute sessions, 4 times per week for 32 weeks (Vadasy et al., 2005). Vaughn, Wanzek, Linan-Thompson and Murray (2007) implemented a successful first grade reading intervention for 30 minutes daily for 10 week repeatable terms. The intervention program studied by Torgesen et al. (1999) required a much smaller weekly time investment of only 80 minutes, however, the duration of the intervention was up to 2.5 years. Based on these results and many other studies, most authorities in the intervention literature recommend intervention terms of 3 to 5 times per week, at least 30 minutes per session, for at least 10 weeks (Burns et al., 2006; Burns & Coolong-Chaffin, 2006; Vaughn & Roberts, 2007; Wanzek & Vaughn, 2007).
Instructor Status

Within a three-tiered model, the role of instructor may be held by a variety of individuals. A trained and licensed teacher, particularly a reading specialist, would logically be the best qualified person to serve as instructor at the Tier 2 level of intervention (Burns et al., 2006). However, the lower cost effectiveness of that approach may limit the availability of the intervention to the school population in need, and the use of instructors with less formal training and cost could be advantageous. In addition, there are arguably some very effective Tier 2 interventions that involve the use of volunteers and peers at little or no additional cost. The Peer Assisted Learning Strategies (PALS; Fuchs et al., 2000) has been shown to deliver significant gains utilizing peer tutors as young as first grade (McMaster et al., 2005). The Help One Student to Succeed program (HOSTS; 1999) utilizes trained adults as instructors, many of whom are volunteers. There is also evidence that this program has helped produce significant reading gains in students K-12 (Burns, Senesac, & Symington, 2004). Other interventions have utilized trained paraprofessionals as the primary instructor in small group or one-to-one intervention ratios and have demonstrated significant reading gains as well (e.g. Jenkins et al., 2004; Torgesen et al., 1999; Vadaszy et al., 2005). Only a minority (i.e. Failure Free Reading, Reading Recovery, LiPS and PASP) of Tier 2 interventions that met the inclusion criteria set by What Works Clearinghouse and Wanzek and Vaughn’s (2007) review of reading interventions required the use of certified or licensed teachers (e.g. Morris et al., 2000).
Summary of Features Common to Evidence-Based Tier 2 Reading Interventions

The selection of one or more interventions to implement at the second tier should be an evidence-based process. There are numerous commercially prepackaged and academically designed interventions available to school-based teams that have demonstrated significant gains in reading for students at Tier 2. Although it is logically desirable to primarily utilize licensed teachers as the primary instructors, it is not necessary as long as licensed teachers and administrators are providing quality training to persons providing instruction (Burns et al., 2006). One-to-one student to instructor ratios have been shown more effective in demonstrating reading gains, yet a ratio of 4 or 6 to 1 also deliver strong gains with the advantage of serving significantly more children at risk. The frequency and duration of evidenced-based interventions in the literature is quite variable, suggesting that significant gains may be demonstrated when Tier 2 interventions are delivered 3-5 times per week for 30 minutes, with duration of at least 10 weeks (Burns et al., 2006; Burns & Coolong-Chaffin, 2006; Vaughn & Roberts, 2007; Wanzek & Vaughn, 2007). Standard components of any Tier 2 reading intervention should be focused on the recommendations of the National Reading Panel (2000). Whatever the intervention selected, it should be implemented with fidelity to the design of the treatment (Burns & Ysseldyke, 2005; Gresham et al., 1993).

Literature Review Related to Propensity Score Analysis Methodology

Selection Bias in Program Evaluations of Tier 2 Intervention Programs

The randomized controlled trial experiment is the “gold standard” for determining causation and therefore determining educational practices that are supported by rigorous
evidence (U.S. Department of Education Institute of Education Sciences, 2003). In a randomized experiment, all of the differences that may exist between groups are equalized through the process of random assignment. This is a control that is effective in balancing group differences whether those differences are observed or unknown, and allows the researcher to regard experimental and control groups as equal (Hahs-Vaughn & Onwuegbuzie, 2006).

Much of the research in the field of education involves comparison of the performances of experimental and control groups whose participants have not been randomly assigned (U.S. Department of Education Institute of Education Sciences, 2003). It is the position of some statisticians that this lack of random assignment prevents researchers from making meaningful conclusions from the results. However, many situations exist in applied settings for which random assignment is not an option. For example, Tier 2 intervention programs typically involve large numbers of participants who have been specifically selected because they have met selection criteria. In applied settings, randomly assigning some students to a control group would deprive them of a service for which they are eligible, which would not be ethical (Shadish et al., 2002). In these situations, referred to as observational studies by researchers, random assignment is neither possible nor desirable, because the intervention is aimed exclusively and intentionally at low achievers. An observational study is an empirical investigation of the effects of an intervention or a treatment when it is not possible to randomly assign subjects to treatment and control groups (Rosenbaum, 2002). Although random assignment is the preferred research technique, researchers have developed methods that
allow them to gain insight into causality when random assignment is not feasible (McNeil, Newman, & Kelly, 1996). The most common methods are the use of a matched control group design, various regression models, and covariance models.

In observational studies, groups are not assigned at random. Treatment groups are assigned based upon some pre-existing criterion rather than by a systematic randomization process. This creates potential bias between the groups in observed and unobserved variables when the effects of treatment are studied (Perkins, Tu, Underhill Zhou, & Murray, 2000). Differences observed in treatment effects may therefore be attributed either to the treatments themselves or to the preexisting bias in the treatment groups (Rosenbaum, 1986).

**Common Methods of Adjusting for Selection Bias in Evaluations of Tier 2 Intervention Programs**

**Regression Models**

The most common methods for adjusting treatment effects in observational studies are regression models (Perkins et al., 2000; Yanovitsky, Zanutto, & Hornik, 2005). In these models, statistical control is gained by attempting to remove any amount of correlation between confounding predictor variables (e.g. socioeconomic status, baseline reading scores) and the outcome variable (e.g. reading gains), with the result that only the unique variation attributable to the treatment variable remains. However, the primary weaknesses of these models are that they often do not identify treatment groups with noncomparable, or non-overlapping, covariates and they assume that covariates are in linear relationship to the criterion variable (Rubin, 1997). For example, selecting
English reading and writing proficiency as separate covariates in predicting success in a Spanish course would overestimate the contribution of variables that likely have substantial overlap, or multicollinearity. Second, invalid conclusions could be drawn if covariate distributions are nonlinear, such as using anxiety as a covariate of test performance. Increasing levels of mild to moderate anxiety are probably positively associated with test performance, while high to very high levels of anxiety are probably negatively associated with test performance (Kirkland, 1971; Shipman & Shipman, 1985).

Another method of attempting to control for selection bias is the instrumental variable approach (Angrist, Imbens, & Rubin, 1996; Heckman, 1995; Winship & Morgan, 1999). In this method, the researcher attempts to identify a set of instrumental variables that have two properties: (a) they have a demonstrated correlation with assignment of subjects to treatment, and (b) they have no correlation with the outcome variable. For example, an instrumental variable affecting assignment to Tier 2 intervention might be class size in a particular school. The treatment effect is estimated after using regression techniques in a three-step process. First, regress the outcome variable on the instrumental variable to estimate a predicted outcome variable (i.e., determine the proportion of unique reading achievement variance attributable to class size). Second, regress the treatment variable on the instrumental variable to determine the predicted treatment assignment for subjects (i.e., determine the proportion of unique variance in intervention group assignment attributable to class size). Third, regress the predicted outcome variable on the predicted treatment assignment for subjects (i.e.,
determine the proportion of unique variance in reading achievement attributable to intervention group assignment). There are multiple difficulties with this technique. For example, it is extremely difficult to identify a variable that meets the definition of an instrumental variable, (i.e., is correlated with treatment assignment but has no correlation with the outcome variable) (Perkins et al., 2000). In addition, the procedure typically produces standard errors that are too large to accurately estimate treatment effects (Winship & Morgan, 1999). Lastly, this procedure is appropriate to utilize only when the treatment effect is known to be uniform for all subjects (Newhouse & McClellan, 1998).

**Analysis of Covariance**

Analysis of Covariance (ANCOVA) is another analytic method commonly utilized by educational researchers when random assignment of subjects is not possible, such as in studies involving comparison of intact classrooms, or when systematic difference in experimental and control groups are evidenced by group differences on a pretest measure (Cohen & Cohen, 1983; Huitema, 1980; Pedhazur, 1997; Stevens, 2007). In this method, a number of variables known to be correlated with the dependent variable are statistically removed from the error variance term. The logic of this method is that if the lack of random assignment to treatment groups is impossible, it is wise to remove the known bias from the equation. Theoretically, this allows the researcher to test the amount of unique variance in the dependent variable accounted for by treatment group membership alone. The method involves first identifying the covariates that may introduce bias in testing differences based on group membership. For example, in a study evaluating the effectiveness of reading interventions, the evaluator may determine that
students assigned to the intervention group are disproportionately male, economically
disadvantaged, African American, and have low oral reading fluency skills. Next, the
magnitude of relationship with the dependent variable and with other potential covariates
is determined through correlational analysis. The covariates for the ANCOVA analysis
are then selected based upon their relative magnitude of correlation with the dependent
variable and their ability to identify unique quantities of the variance relative to the other
covariates. These variances are then statistically removed to reduce systematic bias from
nonrandom group assignment and to reduce the overall within group or error variance
(Stevens, 2007). The result is a comparison of group means that have been adjusted to
“what they would be if all groups started out equally on the covariates” (Stevens, 2007, p.
309). In other words, it provides an estimate of what would have been the result of the
reading intervention if we could eliminate the known effects of low socioeconomic
status, membership in a racial minority group, and being born male?

_Weaknesses of Adjusted Means Approaches Such as Regression and ANCOVA_

One danger in using adjusted means approaches such as multiple linear regression
or ANCOVA to reduce sample bias is assuming that the use of one or many covariates
will statistically equate intact groups (Anderson, 1963; Lord, 1969). This would be
naïve thinking because statistical control can only be utilized for variables that are
measured and tested for relationship to the dependent variable. Any number of untested
variables could also be significant sources of sample bias although unknown to
researchers (Stevens, 2007). Additionally, Huitema (1980) argued that the process of
statistically controlling for one variable could inadvertently attenuate the differences between groups on other variables.

Elashoff (1969) suggested that attempting to equalize the treatment and control groups on a particular covariate may be attempting to create a condition that would not exist in the real world. ANCOVA adjusts the posttest means to what they would be if the groups were equivalent on the covariate, but the groups under study are known to be non-equivalent, which is the reason for utilizing the ANOVA technique. For example, a researcher knows that students in a Title I reading program are not equivalent to their general education peers with regard to reading skills because poor reading skills defined the group eligibility. Adjusting the group means to equalize the effects of pretest group differences does not remove the fact that the groups differ on reading skills. This group will likely have a differential response to the treatment precisely because they have poorer reading skills.

Bryk and Weisberg (1977) described a related weakness in adjusted means comparisons involving growth effects in subjects. They argued that when groups differ on a covariate that is a growth oriented quality, adjusting the means on the basis of the covariate only reduces the pretest bias but not the growth potential bias. For example, adjusting means for differences in oral reading fluency skills in groups may function to equalize the groups, but may not statistically control for the differential response to the reading intervention because groups may still differ in their growth potential for developing increases in that important reading skill. Students with poor reading fluency
skills may be more apt to demonstrate gains than better readers, given the same amount
of practice simply because their low baseline level allows substantial growth potential.

Two statistical assumptions must be satisfied for analysis of covariance
techniques to be appropriate. First, there must be a linear relationship between the
dependent variable and the covariate. If the relationship between the covariate and the
dependent variable is curvilinear, then the adjustment in means will not be accurate or
equivalent for all subjects (e.g., the effect of varying levels of anxiety on test
performance). Second, there must be homogeneity of regression slopes when one
covariate is used or parallel regression planes if more than one covariate is used.
Otherwise, the effect would be similar to an interaction effect (e.g., good readers respond
positively to the treatment but poor readers do not respond). The slope of the relationship
between the dependent variable and the covariate(s) must be similar or parallel in both
the experimental and control groups. (Huitema, 1980; Pedhazur, 1982; Stevens, 2007).

Measurement error in randomized studies serves to detract from the power of the
test, but in nonrandomized studies evaluated with ANCOVA, the effect is to introduce
significant bias in the measured group differences (Pedhazur, 1997). Reichardt (1979)
suggested that since pretest scores are used to identify significant covariates,
measurement error on the pretest can lead to adjusted means that are invalid. In this case,
an estimate of a significant treatment effect could be concluded when none exists.

Type VI Error

A more theoretical criticism of using ANCOVA to analyze group differences in
nonrandomized designs was made by Tracz, Nelson, Newman and Beltran (2005) who
argued that the researcher must be cognizant that the dependent variable in ANCOVA is an adjusted score, and that the testing of differences in this model is based on adjusted (corrected) means. With the effects of each identified covariate statistically factored out of the dependent variable, the remaining variance in the dependent variable is only error variance (Pedhazur, 1997). As a result, this adjusted dependent variable is different than the dependent variable originally measured. When this is done, the adjusted means derived from the ANCOVA may not continue to be measures of the same construct (Tracz et al., 2005). For example, if the construct being addressed in measurement of the dependent variable is reading achievement, the researcher must not assume that adjusted reading achievement and reading achievement are the same things.

A final criticism of using adjusted mean approaches as in ANCOVA is that the researcher is much more likely to commit a Type VI error (Fraas, Newman, Bagakas & Newman, 2006; Newman, Dietchman, Burkholder & Sanders, 1976; Newman, Fraas, Newman, & Brown, 2002; Tracz et al. 2005). Newman et al. (1976) defined a Type VI error as an “inconsistency between the research question and the research methodology” (p. 1). This typically occurs when the researcher applies a common statistical procedure or model to a general research question rather than carefully tailoring the statistical model to a very specific research question. The researcher may be unaware that the selection of a statistical model is actually influencing the formulation of the research question. This is a problem of confusing methods of statistical analysis with experimental design. The result is that technically correct research is produced that is not functionally related to the
true problem. The researcher may draw inaccurate conclusions from the results or make inaccurate interpretations of the data (Newman et al., 1976).

McNeil, Newman, and Kelly (1996) described three ways of examining the unique variance in the dependent variable that can be accounted for by group membership. This unique variance can be examined by logical argument, by the research design employed, and by statistical control. Tailoring the research design can be the strongest way to account for this unique variance, but it is not always possible to include all of the possible confounding variables in the design. Employing statistical control procedures through analysis of covariance can be a powerful alternative when the research design cannot account for all of the unique variation in the dependent variable (Tracz et al., 2005).

Adjusted achievement is not the same thing as achievement. In a study of student achievement outcomes, the construct of student achievement is operationalized by a dependent variable such as scores on the 4th Grade Ohio Achievement Test – Reading (OAT-R). However, if the technique employed to analyze student achievement actually analyzes adjusted scores because covariates were partialed, the analytic technique may not allow the researcher to address the research question directly (Fraas et al., 2006). In other words, when ANCOVA is used to statistically account for covariates that introduce bias in the sample, the resulting construct of adjusted reading achievement should not be confused with the original construct of reading achievement. Adjusted reading achievement is an abstract number. If the researcher is primarily interested in reading
achievement, then there would be a mismatch between the research question and the research methodology (Tracz et al., 2005).

A possible solution to the problem of interpreting adjusted scores appropriately was suggested by Tracz et al. (2005). An adjusted score could be regarded as any other outcome variable that is measured. Its basic measurement properties could be assessed to establish its construct validity. Typically this involves correlating the variable with other measures of the construct that have previously been widely accepted as valid measures (Groth-Marnat, 2003; Huck, 2007; Miller, Linn & Gronlund, 2008; Nunnally, 1978). If the adjusted scores correlate with other accepted measures of the same construct, then they can be interpreted meaningfully. If the adjusted scores do not correlate with other accepted measures of the construct, then they should be regarded as residuals or random error. They can not be interpreted as measures of the construct or anything else until their validity is established (Tracz et al., 2005).

*Propensity Score Analysis: An Alternative Method of Addressing Selection Bias in Program Evaluations of Tier 2 Intervention Programs*

Rosenbaum (2002) and Rosenbaum and Rubin (1983, 1984) developed the propensity score analysis technique as a method of analyzing observational data. The technique is used to analyze the differences between group means when it is not possible to select the treatment groups randomly. Propensity score analysis (PSA) is a way of adjusting the comparison of groups when there is systematic bias between treatment groups on a set of covariates (Tracz et al., 2005). Whereas adjusted means approaches such as multiple regression and ANCOVA try to adjust for bias by making an adjustment
on the dependent variable, propensity score analysis attempts to adjust for this bias by making the adjustment on the independent variable (Yanovitsky, Zanutto, & Hornik, 2005). PSA may be used to make causal inferences from observational studies and to determine if a dataset can be utilized to address causal questions (Rubin, 1997; Perkins et al., 2000; Shadish et al., 2002). It is also useful in eliminating bias that is inherent when comparing treatment groups in observational studies. Fundamentally, propensity score analysis is a measure of differences between group means (D’Agostino, 1998).

There are several advantages to using propensity score analysis over multiple regression or ANCOVA. First, treatment effects can often be estimated without specifically developing a model of the relationships between the covariates and the outcome variable. Instead, the propensity score analysis begins by statistically addressing the selection bias problem rather than measuring the outcome variable’s correlation with individual covariates (D’Agostino, 1998). Second, modeling of the propensity score is less vulnerable than multiple regression to possible effects of including variables in the wrong functional form, such as an adjusted outcome variable (Drake, 1993; Perkins et al., 2000; Rubin, 1997). Third, since propensity score analysis emphasizes modeling the selection bias problem independent of the outcome variable, the result is a balancing of subject selection that approximates random assignment to treatment groups (Rosenbaum & Rubin, 1983). Fourth, the propensity score can be used for adjusting selection bias for more than one outcome variable. Regression techniques require that the set of covariates is modeled on each outcome variable (Yanovitsky et al., 2005).
Propensity score analysis accounts for differences between treatment and control groups in a fundamentally different way than ANCOVA. Whereas ANCOVA attempts to remove the effects of covariates from the measurement of the outcome variable, propensity score analysis accounts for the differences between treatment and control groups by modeling the process of group selection (McCaffrey et al., 2004). Rosenbaum and Rubin (1983) defined the propensity score as “the conditional probability of assignment to a particular treatment given a vector of observed covariates” (p. 41). They showed that observed covariates that may represent sources of group selection bias are independent of group assignment conditional to the propensity score. An individual subject’s propensity score is the subject’s probability of being assigned to the treatment group (or a level of treatment, depending upon the structure of the variable), as a condition of the individual subject’s covariate measures (Rosenbaum & Rubin, 1983). In studies or program evaluations involving large numbers of subjects, the determination of the propensity score for each subject allows the researcher to observe the equal distribution of covariates within both groups with the result that the covariates do not confound the measurement of treatment effects (Rosenbaum & Rubin, 1983). The propensity score becomes a balancing score. Subclasses of subjects are formed based on the subjects’ similar propensities for assignment to treatment and their covariate values, with the result that the distribution of the covariates across treatment groups is equalized. Applied to observational studies, Rosenbaum and Rubin (1984) later suggested that the propensity score can be used to adjust for pretreatment bias by using matched sampling and stratification techniques. Other methods of adjusting for pretreatment bias using
propensity score modeling have been suggested that involve applying weighted values to observations (Farley et al., 2003; Hirano, Imbens, & Ridder, 2003).

Literature Review Related to Applications of Propensity Score Analysis

Use of Propensity Score Analysis Methodology in Applied Settings

The propensity score analysis methodology fills a particular research niche that is common in applied settings. There are many instances in which it is not possible to conduct experimental research to determine treatment effects. In these cases, researchers have used quasi-experimental techniques to accumulate persuasive evidence (Shadish et al., 2002; West, Biesanz, & Pitts, 2000). For example, random assignment to treatment conditions is often not possible in studies of the effectiveness of medical interventions. It is often the case that patients choose to participate in interventions based upon personal characteristics that are inherent sources of bias in studying treatment effects (McCaffrey, Ridgeway, & Morral, 2004). In addition, even when random assignment is possible, researchers frequently face ethical dilemmas in using random assignment in community treatment studies. When conventional wisdom favors one treatment condition, researchers have a difficult dilemma in assigning some patients to an alternative placebo condition (McCaffrey et al., 2004; Shadish et al., 2002). Likewise, educational researchers and practitioners are faced with many situations in which they must evaluate the effectiveness of programs when it is impossible to randomly assign subjects to treatments (e.g., when specific eligibility criteria are imposed on group identification by legislative mandates). It is also often unethical to withhold treatment (e.g., special education services, academic or behavioral interventions) students are eligible to receive.
Early Studies Utilizing Propensity Score Analysis in the General Scientific Literature

The propensity score analysis technique emerged from the field of human medicine where researchers in the field are often limited to observational studies because of ethical concerns about assignment of subjects to treatment groups (D’Agostino, 1998). The originators of the technique were faculty in the departments of Statistics and Human Oncology at the Universities of Wisconsin, Madison, and the University of Chicago, respectively (Rosenbaum & Rubin, 1983). Early applications of the technique in the medical field included studies comparing triple-bypass cardiac surgery to medication treatment (Myers et al., 1987), studies of the cardiac medication Digoxin’s effectiveness compared to standard medications (Muller et al., 1986), studies of survival rates of patients who received triple-bypass surgery early in treatment versus later in treatment (Myers et al., 1987), studies of myocardial infarct patients admitted to differing levels of hospitalization (Fiebach et al., 1990), studies comparing survival rates of cardiac patients diagnosed specifically with triple-artery disease versus severe angina pectoris (Myers et al., 1989), studies comparing the effects of hospitalized and ambulatory patients with community-acquired pneumonia (Stone, Obrosky, Singer Kapoor & Fine, 1995), and studies investigating the prevalence of caesarean procedures among nulliparous women who received epidural analgesia versus those who did not (Lieberman et al., 1996).

Early Applications of Propensity Score Analysis in Mental Health and Social Science Literature

A small number of applications of propensity score analysis appeared in the mental health literature in the 1980’s and 1990’s. For example, Lavori, Keller and
Endicott (1988) used this method of analysis when studying the validity of a diagnostic instrument to diagnose major affective disorder in uninterviewed family members of already diagnosed patients. A second study utilizing this statistical method investigated the validity of including or excluding data from patients in psychopharmacological clinical trials when the participants deviated partially from the standard pharmacological protocol (Lavori, 1992).

Propensity score analysis appeared in the social sciences literature in a limited way during this time period as well. For example, Berk and Newton (1985) used the technique when comparing the rates of recurrence of spousal abuse among perpetrators who had been arrested versus those who had not been arrested by law enforcement. In another study involving the results of spousal abuse, researchers used propensity score analysis to determine the effects of the timing of seeking refuge in a shelter for battered women (Berk, Newton, & Berk, 1986). Other early applications of propensity score analysis in the social science literature involved projecting income statistics and tax revenues in local governments (Czajka, Hirabayashi, Little, & Rubin, 1992), and comparing academic achievement growth in public versus parochial schools (Hoffer, Greeley, & Coleman, 1985).

Contemporary Applications of Propensity Score Analysis in the General Scientific Literature

In order to determine the contemporary application of propensity score analysis in the general literature, the Elton Bryson Stephens Company (EBSCO) Host Research Databases-Academic Search Premier was used on 4-8-08 to search for references to
propensity score analysis in the academic literature. A search of the words “propensity score” in any combination within the text or abstract of any available publications yielded 445 related studies. In order to identify more contemporary applications of the analysis technique, this search was further refined by maintaining the above parameters and limiting the year of publication to 2001 or later. This parameter limited the available related studies to 428.

Because propensity score analysis was developed to analyze the results of observational studies in the medical field, it was important to determine the extent to which the contemporary applications of the technique remain medically oriented. In order to determine this, the search was further refined by maintaining the same search parameters but restricting the sources of publication to medical journals. This set of restrictions resulted 310 matches out of the 428 citations overall. Between January, 2001 and April, 2008, of all citations of peer reviewed research applying propensity score analysis, 72.4% were published in medical journals. Clearly, the technique was being utilized in observational studies in which causal inferences were desired, but the nature of the research did not allow random assignment of groups either for ethical reasons or because of non-random occurrence of medical conditions in the population. The remaining 118 articles published between January, 2001 and April 2008 could be classified as pertaining to economics and social sciences (50); measurement and statistical applications (24); criminal justice applications (13); psychology and child development (12); substance abuse literature (10); and education literature (9).
Contemporary Applications of Propensity Score Analysis in the Social Sciences

Literature

Propensity score analysis has been used to study the effects of social welfare policies in applications where random assignment is not possible because specific qualifying criteria apply to eligibility for welfare benefits. For example, propensity scores were used to study the effects of the National Welfare to Work strategies on the cash benefits of sanctioned and non-sanctioned welfare recipients (Peck, 2007); the effects of the 1995-1998 Ontario welfare reforms on the utilization of welfare resources by single mothers (Brzozowski, 2007); the effect on several sub-groups of the New Hope project, a program providing welfare program benefits to working families with incomes below the poverty level (Gibson, 2003); the effects of the Minnesota Family Investment Program policies on income and employment of hard-to-employ welfare recipients (Yoshikawa, Magnuson, Bos, & Hsueh, 2003); to evaluate the effect of food stamps on the perception of welfare recipients of food insecurity (Gibson-Davis & Foster, 2006); to evaluate the extent to which childbearing affected the material well-being of female welfare recipients (Aassve, Mazzuco, & Mencarini, 2005); and to evaluate the effect of the Massachusetts workforce development system on the earnings of several sub-groups of economically disadvantaged adults (Raphael & Stoll, 2006).

Several studies in the social science literature have addressed the politics of environmental policies and have incorporated propensity score analysis as part of the methodology to address the observational nature of their data. Fredriksson and Millimet (2004) used propensity score matching to determine that presidential-congressional
political systems set significantly lower taxes on environmental pollution than parliamentary regimes. Propensity score matching was also used to determine if the requirements of the 1970 Clean Air Act were the cause of a subsequent 80% reduction in sulfur dioxide air pollution in the U.S. (Greenstone, 2004). Jimenez (2005) applied propensity score stratification to the analysis of data related to the amount of technological innovation resulting from voluntary and non-voluntary environmental regulations. In a follow-up study of environmental regulations in Chile, Jimenez (2007) furthermore used propensity score matching to determine that voluntary environmental agreements between corporations and political entities resulted in incremental pollution reductions that also preserved the corporations’ financial viability.

The social science literature also contains many applications of propensity score analysis to the study of several features of the political process. For example, Forsman and Isaksson (2003) compared the effectiveness of web-based opinion polls and traditional telephone opinion polls in determining voter intentions using propensity score matching. Lee (2004) used propensity scores to adjust the weighting of responses to web-based surveys based on demographic stratification. The effect of types of voting technologies on residual votes was studied by Bailey (2006) utilizing both regression models and propensity score matching. Propensity score matching was also used to determine if voter mobilization phone calling banks were effective in increasing voter turnout compared to elections in which the strategy was not used (Arceneaux, Gerber, & Green, 2004). Barabas (2004) used propensity score analysis to determine whether political discussion and deliberation affect the formation of political opinions in voters
with differing degrees of opinion strength on an issue. In another study of voter opinions of candidates’ perceived honesty, Cassino (2007) used propensity score matching to study the effects of verbalizing opinions of the candidates on accurate perceptions of honesty. The research concluded that verbalizing opinions decreased voters’ accuracy of honesty judgments about the candidates. In another observational study utilizing propensity score analysis, Gallagher (2006) studied the effects of candidates’ gender on the gender distribution in voter turnout, and concluded that the presence of a female candidate increased female voter turnout but did not influence male voter turnout.

*Contemporary Applications of Propensity Score Analysis in the Criminal Justice Literature*

Because propensity score analysis was designed to address observational research in which random assignment to treatment is not possible, it has been useful to researchers in the field of criminal justice to study the impact of covariates such as incapacitation, marital status, age and race. It was used by Ridgeway (2006) to analyze data from traffic stops to determine the extent to which race bias affects the duration of stops, the search rates, and the rates of receiving traffic citations. Berk, Li, and Hickman (2005) re-analyzed data from Maryland capital cases using propensity score analysis to determine the impact of race in sentencing decisions. Their results, using a more appropriate method of analysis, suggested that the impact of race reported in earlier traditional studies is not as significant as previously reported. The effect of age of onset of substance abuse disorders on later criminal activity was investigated by Slade et al. (2008). They used propensity score matching to estimate the effects of substance abuse disorders that began
before age 16 and after age 16, and determined that the subjects whose substance abuse began before age 16 had a rate of incarceration four times higher. Propensity score matching was used to compare the effects of an intensive treatment program and the usual lock-up treatment for juvenile boys (Caldwell, Vitacco, & van Rybroek, 2006). Use of the technique revealed that intensive treatment reduced the likelihood of recidivism from 44% to 23%.

The effect of marriage on crime in early adulthood was the subject of research applying propensity score analysis by King, Massoglia, and Macmillan (2007). They used propensity score matching to determine if marriage affects criminal activity, whether results are consistent for both sexes, and whether the propensity to marry affects the deterrent effects of marriage. They concluded that marriage has a significant suppressor effect on criminal activity for males even when considering their likelihood to marry. In addition, they concluded that males who are least likely to marry showed the greatest reduction in criminal likelihood. The preventive effects of incapacitation were estimated using propensity score matching (Sweeten & Apel, 2007). Incarcerated and non-incarcerated felons were matched on an extensive number of confounding covariates. Sweeten and Apel concluded that a range of 6.2 to 14.1 crimes are prevented per year of incarceration for juveniles, and 4.9 to 8.4 crimes are prevented per year of adult incarceration.
Contemporary Applications of Propensity Score Analysis in the Psychology and Child Development Literature

The fields of child development and developmental psychology have frequently utilized propensity score analysis to analyze data that must be gathered observationally, such as from children placed in foster care, from children whose mothers are employed at various degrees outside the home, and family poverty data. For example, the effects of maternal employment on the development of three-year-olds’ vocabulary and behavioral adjustment was studied in black, white and Hispanic children utilizing propensity score matching models to address the covariates of race when maternal employment could not be randomly assigned (Berger, Brooks-Gunn, Paxson, & Waldfogel, 2008; Hill, Waldfogel, Brook-gunn & Wen-Jui, 2005). The effects of varying levels of economic disadvantage on the emotional, social and academic adjustment of children were studied although the outcome data were almost certainly biased by a large number of covariates and selection biases. Votruba-Drzal (2006) used propensity score analysis to address the methodological challenges of this income study among children in middle childhood.

Youth who have chosen to enter gang membership represent another clearly non-random population. Haviland, Nagin, Rosenbaum and Tremblay (2008) utilized propensity scores to balance observed covariates in their investigation of the impact of gang membership on incidents of violent delinquency.

Children placed in foster care represent another research scenario in which subjects cannot be randomly assigned to treatment. Bellamy (2008) utilized propensity score matching to determine the level of well-being of children who were reunified with
their families after a foster care placements of 8 months or more compared to similar children who could not be reunified with their families. In a similar study, the effects of placement stability on the behavioral outcomes of children in the foster care system were studied by Rubin, O’Reilly, Xianqun, and Localio (2007). They used propensity scores to match children on level of placement stability and determined that the children achieving a stable foster placement within 45 days had significantly fewer behavioral problems than children whose foster placements became stable after 45 days or never reached stability. Propensity score matching was used to compare the adjustment of children placed in traditional foster care with those placed in the SAFE Homes program, a short-term group care program for children who are in a first protective care placement (DeSena et al., 2005). Use of the matching technique allowed DeSena et al. to determine that traditional foster care placement resulted in more favorable adjustment of children placed in protective care at a significantly lower cost.

The child development literature includes a small number of program evaluations investigating the effects of various instructional and mental health interventions on the adjustment of children targeted by the programs. Because the interventions employ selection criteria and are not delivered to a random sample of subjects, selection bias is inherent in the research designs. A structured arts program was targeted at adolescents from low-income communities and was designed to improve psychosocial outcomes as well as artistic ability (Wright et al., 2006). Comparisons with a control group assigned using propensity scores indicated that the structured arts program was associated with a reduction in emotional problems in low income adolescents. The mental health outcomes
of children were compared for children who received services from a system-of-care program and those receiving services from traditional care systems (Foster, Stephens, Krivelyova, & Gamfí, 2007). Propensity score matching methodology was used to analyze differences in clinical and functional outcomes. Results showed that children in system-of-care programs showed significantly better mental health improvements than their counterparts.

Contemporary Applications of Propensity Score Analysis in the Substance Abuse Literature

Propensity score analysis has been the most common statistical method applied to evaluating the effectiveness of public health interventions when nonrandom assignment conditions exist (McCaffrey et al., 2004). Evaluating the effectiveness of community-based substance abuse treatments is a prominent example often cited in descriptions of propensity score analysis applications because the effectiveness of these programs must be evaluated despite the complex set of factors contributing to the selection of subjects. For example, propensity score analysis has been applied to studies evaluating the effectiveness of the National Youth Anti-Drug Media Campaign (NYAMC) (Hornik, et al., 2002; Yanovitsky et al., 2005), to the effectiveness of treatment programs targeted at adolescents and young adults in major U.S. cities (Gerstein & Johnson, 1999; Hser, et al., 2001), to the effectiveness of the Drug Abuse Reporting Program (DARP) (Sells & Simpson, 1979), and to evaluation of the National Institute on Drug Abuse (NIDA) Treatment Outcome Prospective Study (TOPS) (Hubbard, et al., 1985). The technique has also been applied to studying the effects of tobacco retailer density in neighborhoods
and youth cigarette smoking (Novak, Reardon, Raudenbush & Buka, 2006), the differential response to cognitive behavioral therapy among youth with methamphetamine and cocaine abuse disorders (Callaghan, Taylor, Victor, & Lentz, 2007), the relationship between the use of hard drugs in high school and adult occupational outcomes (Ringel, Ellickson, & Collins, 2007) the relationship between juvenile delinquency and adult marijuana and cocaine use (Doherty, Green, & Ensminger, 2008), the effects of parental substance abuse treatment on the incidence of repeated reports of child abuse (Guo, Barth, & Gibbons, 2006), and the rates of criminal arrest for opiate users participating in substance abuse treatment (Campbell, Deck, & Krupski, 2007).

Contemporary Applications of Propensity Score Analysis in the Education Literature

Between January, 2001 and May, 2008, eleven articles appeared in the education literature in which propensity scores were used to address confounding covariates in observational studies. Within the school psychology literature in particular, Wu, West, and Hughes (2008) investigated the effects of kindergarten retention on reading and math achievement of students over a three-year follow up period. They used propensity scores to match retained and unretained students in order to determine the effects of treatment. Treatment effects were significant and negative in mathematics and insignificant in reading. This application of the technique was based on the consideration that retention is an individually determined intervention and random assignment to the intervention would be unethical, a decision analogous to assignment to a Tier 2 reading intervention. A similar study investigated the effect of kindergarten retention on children’s socio-
emotional development (Hong & Yu, 2008). Researchers used propensity score stratification in attempt to address the selection bias inherent in decisions to use grade retention. They concluded that effect size attributed to the intervention was insignificant and that kindergarten grade retention does not negatively affect socio-emotional development. In another study involving early-childhood interventions, selection bias in preschool enrollment was also addressed in a study investigating the academic effects of preschool experience through fourth grade.

Hill (2006) compared the achievement of 4th graders with and without preschool experience by using propensity scores to address the selection bias inherent in the family’s decision to enroll four-year-olds in preschool programs. One large-scale program evaluation involving comparisons of developmental gains made by children attending federal Head Start programs and state sponsored early childhood programs was conducted by Henry, Gordon, and Rickman (2006). Henry et al. used propensity score techniques to match preschool students attending Head Start with children otherwise eligible who attended a state sponsored preschool program in Georgia. The matching technique allowed for comparison of self-selected groups. Although statistically similar at the beginning of preschool, the students attending state sponsored preschool made greater gains in 5 of 6 developmental domains assessed. In each of these studies, the assignment to an educational intervention could not be conducted at random (i.e., retention, Head Start) and that selection bias was addressed in the evaluation of intervention effectiveness by utilizing propensity score analysis. These program
evaluation scenarios are similar to the current problem affecting nonrandom assignment to Tier 2 academic interventions within a response-to-intervention framework.

Several studies involving specialized populations of adolescents and specialized interventions appear in the literature applying propensity score analysis to address selection bias. Callahan, Wilkinson, and Muller (2008) investigated the effects of the local concentration of ESL students in high schools and the degree of generational separation from immigrant status on academic achievement. They used propensity score matching to make comparisons between students attending schools with few English language learners (ELL) and those attending high-concentration schools. Results showed that ELL placement had a positive effect among second-generation Spanish speakers in high concentration schools, but may have a negative effect on first generation Spanish speakers in low-concentration schools. Wolf and Wolf (2008) evaluated the effectiveness of an alternative school transition program for adolescents to determine if the program improved student attendance or decreased the likelihood of further alternative school placement. They used propensity score matching to identify matched pairs of students to determine any differential effects between those participating in the transition program, reporting significant positive treatment effects for attendance but significant negative treatment effects for reassignment to alternative school.

Propensity score analysis was recently applied to the debated association between extensive work hours and high school dropout. Previous research suggested that high school students who worked extensive hours in addition to attending high school were significantly more likely to drop out of high school than their peers who worked few

hours. However, the research was complicated by preexisting differences on a number of personal variables. This problem was addressed in a study by Lee and Staff (2007) employing propensity score matching. When participants were matched on propensity scores that balanced treatment groups based on socioeconomic background, school performance, personal aspirations and orientation toward work, the researchers concluded that working extensive hours on a job does not encourage students to drop out of high school.

Systematic differences between students who choose to take advanced courses in mathematics and science and those who choose general courses have historically confounded the research on the question of whether taking advanced courses improves scores on basic achievement tests. An application of propensity score analysis techniques by Leow, Marcus, Zanutto, and Boruch (2004), addressed several known covariates that predisposed some students to take advanced courses. After balancing students on the basis of their propensity scores, Leow et al. found significant positive treatment effects for advanced coursework upon basic achievement test scores in mathematics and science.

Furstenberg and Neumark (2007) analyzed data from the Philadelphia Educational Longitudinal Study (PELS) to determine if a set of programs designed to increase high school graduation and college attendance were effective. They utilized propensity score matching techniques to attempt to correct for selective participation in the programs since the interventions were voluntary. After balancing the treatment and control groups based on their propensity scores, Furstenberg and Neumark concluded that
these programs increased the likelihood of high school graduation and positive attitudes about college attendance.

One study utilizing a propensity score analysis technique was reported in the education literature addressing the development of reading skills in the early grades (Condron, 2008). This study investigated the effects of skill-based grouping in the delivery of the core curriculum in grades K-3 compared to students instructed heterogeneously. Students placed in low, middle, and high reading groups were matched with similar peers based on their propensity scores on several observed covariates. The results indicated that students in skill-based reading groups had unequal reading gains compared to matched peers.

Dual Analyses: Comparisons of Regression Techniques and Propensity Score Analysis in Educational Research

As the application of propensity score analysis becomes more prevalent in observational studies, a small number of studies have attempted to determine if propensity score analysis is a more appropriate method of addressing selection bias and also illustrate that utilizing the method can produce different conclusions than multiple regression analysis when applied to educational research. For example, Titus (2007) addressed the problem of self-selection bias in the financial results of earning a Master’s degree in several different academic disciplines. By matching subjects based on propensity scores, Titus was able to quantify significant mean treatment effects of earning a Master’s degree on financial earnings. This conclusion was not demonstrated when utilizing conventional multiple regression techniques.
Frisco, Muller, and Frank (2007) used both propensity score matching and ordinary least squares regression methods to compare the mathematics achievement of adolescents whose parents’ marriages were intact and those whose parents’ marriages were dissolving. They attempted to address the methodological challenge of potential selection bias by analyzing the data twice, using both a least squares regression method and propensity score matching. Using data from the National Longitudinal Study of Adolescent Health and the Academic Achievement Study, they compared the achievement of groups over a one-year period. Their results indicated significant association between adolescent mathematics achievement and the status of the parental marriage union when either least squares regression or propensity score analysis was utilized.

An internal program evaluation conducted in the Montgomery County Public Schools in Rockville, Maryland sought to determine if a professional development program for teachers resulted in improved reading and mathematics achievement for students (Modaressi & Wolanin, 2007). The program evaluators utilized both analysis of covariance (ANCOVA) and propensity score stratification techniques to determine if achievement improved after considering the effects of a number of student covariates such as English Language Learner (ELL) status, economic disadvantage, and students’ prior achievement. Results indicated no significant effect of the professional development regardless of the data analysis method used.

In a similar study involving hypothetical data, Fraas, Newman, Bagakas, and Newman, (2006) simulated an educational application comparing the achievement
outcomes of students in a data set when the criterion variable was influenced by a number of covariates. They addressed the significant covariates using two different methods and compared the power of the analyses to detect treatment effects. Fraas et al. demonstrated that utilizing propensity score analysis by stratification of propensity scores yielded significant treatment effects while analysis of the data using conventional Analysis of Covariance methods failed to detect significant treatment effects.

*Propensity Score Analysis Applied to Reading Program Evaluation*

The effectiveness of Title 1 reading programs in the Lansing, Michigan school district was evaluated to determine if schools implementing the Reading First program had higher reading gains than schools not using Reading First (Carlisle, Schilling, Zeng, Cortina, & Kleyman, 2006). The unit of comparison was the entire population of each selected school building. In other words, all students in some schools received Reading First, while no students in other schools received the intervention. Differing rates of poverty and ethnic makeup between schools were known to be confounding variables in the study. The program evaluators utilized propensity score analysis as the preferred method to evaluate the differences between group gains while accounting for the differences in the confounding covariates. Propensity modeling was conducted using the matching technique, and schools were grouped into six matched sets of Reading First and non-Reading First schools. A comparative analysis using regression models or ANCOVA was not conducted. Results indicated that students in Reading First schools had higher achievement gains in several areas of reading skill assessed (Carlisle et al., 2006).
The Madison Municipal School District (Madison, Wisconsin) utilized propensity score matching in its internal evaluation of the effectiveness of the Reading Recovery program (Potter, 2004). Researchers compared the longitudinal reading achievement data of students who qualified for Reading Recovery to those who did not, and used propensity score analysis to address the selection bias inherent in the eligibility criteria. Observed covariates of special education status, race, educational level of parents, single parent household, prior achievement, and economic disadvantage were balanced on the propensity score. Effect sizes were determined based only on the propensity analysis and there was no comparison of effect size estimates obtained through ANCOVA. Results showed modest longitudinal gains for students who reached the status of “discontinued” from eligibility for the Reading Recovery program, but no significant improvements for students who had participated in Reading Recovery regardless of length of time in the program.

**Gap in the Educational Literature Filled by the Proposed Study**

As this review of literature illustrates, the applications of propensity score analysis remain largely within the fields of medical research, with 72.4% of articles appearing between 2001 and 2008 in medical journals. Outside of the medical literature, the technique has been applied most frequently in the social science literature. The smallest percentage of studies applying propensity score analysis has appeared in the education literature. However, the research conditions that make the technique useful, such as lack of random selection, the presence of observed covariates, and a desire on the part of researchers to make causal inferences make the application to many research
situations in education appropriate. In particular, local program evaluations of the effectiveness of Tier 2 interventions fit these conditions. Students are assigned to these interventions because they meet particular eligibility criteria, making the selected sample nonrandom. Students requiring this level of intervention are disproportionately African American, from economically disadvantaged backgrounds, and English language learners.

Program evaluation of large-scale Tier-2 intervention programs has traditionally been conducted through two methods. The preferred method is the randomized controlled trial experiment, used when subjects can be randomly assigned to experimental and control groups (U.S. Dept. of Education Institute of Education Sciences, 2003). A less preferred option is a quasi-experimental study that attempt to control for confounding covariates by either matching subjects or using adjusted means approaches to gain statistical control of known covariates (i.e., Analysis of Covariance (ANCOVA) or regression techniques). Propensity score analysis has been utilized to evaluate the effectiveness of Tier 2 intervention programs on a very limited basis and under limited circumstances (e.g., Carlisle et al., 2006; Potter, 2004). The studies conducted thus far have occurred in large urban school districts in which the unit of comparison has been selected at the school level rather than at the student level (e.g., Carlisle et al.), or students have been matched on the propensity score to narrow the student population under study (e.g., Potter). Neither study investigated the possibility that effect sizes determined by more traditional data analysis methods (i.e., multiple regression analysis or
Analysis of Covariance) might differ when data are analyzed via propensity score analysis (see Fraas et al., 2006).

The current study proposes to address this gap in the literature by conducting a local program evaluation of a Tier 2 reading intervention program in which students are selected for study based upon intervention eligibility criteria, specifically, students whose oral reading fluency scores are among the lowest 20 percent for the class. Propensity score stratification will be utilized to balance subjects on known covariates (e.g., socioeconomic status, racial minority group membership, oral reading fluency skills, gender). Propensity score stratification is a method of controlling for selection bias by making an adjustment on the independent variable (i.e., assignment to treatment or non-treatment condition) by grouping subjects into strata determined by observed characteristics on the known covariates (Hodges & Grunwald, 2005; Rosenbaum & Rubin, 1983). After the strata are defined, treatment and control subjects in each stratum can be compared directly. Stratification allows for unequal sized groups in treatment and control conditions at each stratum. An alternative approach, propensity score matching, uses the propensity score to identify a matched control subject for each subject in the treatment condition. When an ample number of control (nontreatment) subjects are readily available, stratification is the preferred method because it utilizes more of the available data and can better account for the differences in group sizes between treatment and control groups (D’Agostino, 1998; Hahs-Vaughn & Onwuegbuzie, 2006).

The research questions will be addressed by treating the data both by Analysis of Covariance and propensity score analysis, and a comparison of the research conclusions
driven by each method will be made. The value of the study will be to determine if
analysis of the data by ANCOVA and propensity score analysis approaches identify
similar or different effect sizes for the Tier 2 intervention. These results will inform a
discussion of advantages and disadvantages of the two methods that school psychologists
may utilize to conduct local program evaluations of Tier 2 intervention programs.

**Literature Review Related to the Selected Covariates in Propensity Analysis and
ANCOVA Procedures**

The use of propensity score analysis requires the selection of a set of covariates
on the basis of empirical evidence from prior research rather than identifying
confounding variables from the actual data used in the study. It is critically important to
include only potentially confounding variables as covariates. Identifying covariates from
the study data may identify variables that measure some effect of the treatment itself
(Yanovitsky et al., 2005). In this section, literature will be reviewed for several
demographic and measurement variables that have been well established in the education
literature as covariates with reading achievement. Literature will be reviewed
demonstrating that the demographic variables of race, gender, socioeconomic status, and
English Language Learner (ELL) status have repeatedly been observed as covariates
with student reading achievement. In addition, reading fluency is discussed as a
significant predictor of overall reading achievement.

**Race**

One of the key provisions of the No Child Left Behind Act (NCLB) of 2001 was
the setting of absolute achievement targets for all students, called adequate yearly
progress (AYP). Determinations of AYP are required for all students as an aggregate and also for key subgroups of students identified by race, economic disadvantage, disability, and English language learner status (U.S. Department of Education Office of Elementary and Secondary Education, 2002). The requirement to disaggregate data based on these demographic variables is based on the substantial body of data identifying significant achievement gaps between majority white middle class English speaking students and their minority counterparts (U.S. Department of Education Institute of Education Sciences, 2006). These achievement gaps exist in both reading and mathematics.

The National Council for Educational Statistics’ National Assessment of Educational Progress (NAEP) has conducted nationally representative standardized assessments of U.S. students in grades 4 and 8 for over 30 years. Data from the NAEP have consistently indicated significant achievement gaps between white students and Hispanic and black students at both grade levels (U.S. Department of Education Institute of Education Sciences, 2005) (see Table 4).

Studies of the reading achievement levels among African-American and white students in Ohio schools indicate that a significant achievement gap exists between these populations (State Board of Education Closing Achievement Gaps Task Force, 2003). For example, among Ohio’s sixth-graders in 2003, although two-thirds of white students met reading proficiency standards, only one-fourth of African-Americans were proficient. In that year, only 11% of African-American students passed all sections of the Ohio Proficiency Test while 43% of white students passed all sections. White students are much more likely to be identified as gifted than their African-American counterparts, and
African-American students are much more likely to be served in special education programs than white students (State Board of Education Closing Achievement Gaps Task Force, 2003).

Table 4

Grades 4 and 8 Mean Reading Scaled Scores for Available School Years 1998-2005 by Race

<table>
<thead>
<tr>
<th>Year</th>
<th>white</th>
<th>black</th>
<th>achievement gap</th>
<th>Hispanic</th>
<th>achievement gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>226</td>
<td>193</td>
<td>33*</td>
<td>193</td>
<td>33*</td>
</tr>
<tr>
<td>2000</td>
<td>224</td>
<td>190</td>
<td>34*</td>
<td>190</td>
<td>34*</td>
</tr>
<tr>
<td>2002</td>
<td>229</td>
<td>199</td>
<td>30*</td>
<td>201</td>
<td>28*</td>
</tr>
<tr>
<td>2003</td>
<td>229</td>
<td>198</td>
<td>31*</td>
<td>200</td>
<td>29*</td>
</tr>
<tr>
<td>2005</td>
<td>229</td>
<td>200</td>
<td>29*</td>
<td>203</td>
<td>26*</td>
</tr>
<tr>
<td>Grade 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>270</td>
<td>244</td>
<td>26*</td>
<td>243</td>
<td>27*</td>
</tr>
<tr>
<td>2002</td>
<td>272</td>
<td>245</td>
<td>27*</td>
<td>247</td>
<td>25*</td>
</tr>
<tr>
<td>2003</td>
<td>272</td>
<td>244</td>
<td>28*</td>
<td>245</td>
<td>27*</td>
</tr>
<tr>
<td>2005</td>
<td>271</td>
<td>243</td>
<td>28*</td>
<td>246</td>
<td>25*</td>
</tr>
</tbody>
</table>


*p < .05
Economic Disadvantage

There is a broad and longstanding literature describing the relationship between socioeconomic status and reading achievement. The social science literature contains descriptive studies describing the impact of socioeconomic status on the achievement of nations. For example, the gross domestic product (GDP) per capita correlates significantly with mean reading achievement by country (Baker, Goesling, & Letendre, 2002; Heyneman & Loxley, 1983). This may be the result of increasing levels of government spending on education in higher GDP countries, better nutritional standards, better healthcare, or better knowledge of potentially debilitating environmental factors (e.g. lead poisoning) in wealthier countries (UNICEF, 2008; Vernon-Feagans, Hammer, Miccio, & Manlove, 2001).

The economic condition of the school a child attends also is predictive of academic achievement (Ogle et al., 2003). Reading achievement among students attending private schools in the U.S. is higher than for students attending public schools (Snow, Burns, & Griffin, 1998). The financial resources available to private schools are generally higher than for public schools, and private school students have greater access to educational materials than their public school counterparts. Differences in academic performance of students tend to be significantly greater between schools in the U.S. than within schools, with private school students outperforming public school students (Ogle et al., 2003).

Research has also shown that the access to financial resources at the family level is significantly correlated with academic achievement (Bradley & Corwyn, 2002;
Brooks-Gunn et al., 1999; Chiu & McBride-Chang, 2006). Higher socioeconomic status has implications for the amount of financial capital of families as well as other types of resources, such as education level of parents and level of social connections available such as through parental job status (Chiu & McBride-Chang, 2006). At the family level, higher levels of SES can affect student achievement through higher parental expectations (Battin-Pearson et al., 2000), more effective parenting skills (Bornstein & Bradley, 2003) and through broader learning opportunities (Benabou, 1996).

Studies of the reading achievement levels among economically disadvantaged students and the general population of students in Ohio schools indicate that a significant achievement gap exists between these populations (State Board of Education Closing Achievement Gaps Task Force, 2003). For example, at the elementary school level there is a significant relationship between the percentage of students attending the school whose family income is below the poverty level and reading achievement scores. For every 10% increase in a school’s poverty level, the Ohio Proficiency Test reading achievement scores decrease by 5.6 percentage points.

The National Council for Educational Statistics’ National Assessment of Educational Progress (NAEP, 2005) has conducted nationally representative standardized assessments of U.S. students in grades 4 and 8 for over 30 years. One indicator of a student’s socioeconomic status is eligibility for free or reduced lunch as part of the National School Lunch Program. Data from the NAEP have indicated a significant achievement gap between students who are eligible for the program since eligibility data have been collected (see Table 5).
Table 5

*Grades 4 and 8 Mean Reading Scaled Scores by Student Eligibility for Free/reduced Lunch for Available Years 1998-2005*

<table>
<thead>
<tr>
<th>Year</th>
<th>scaled score (free/reduced)</th>
<th>scaled score (non-eligible)</th>
<th>achievement gap</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grade 4</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>196</td>
<td>223</td>
<td>27*</td>
</tr>
<tr>
<td>2000</td>
<td>193</td>
<td>225</td>
<td>32*</td>
</tr>
<tr>
<td>2002</td>
<td>203</td>
<td>230</td>
<td>27*</td>
</tr>
<tr>
<td>2003</td>
<td>201</td>
<td>229</td>
<td>28*</td>
</tr>
<tr>
<td>2005</td>
<td>203</td>
<td>230</td>
<td>27*</td>
</tr>
<tr>
<td><strong>Grade 8</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>245</td>
<td>269</td>
<td>24*</td>
</tr>
<tr>
<td>2002</td>
<td>249</td>
<td>272</td>
<td>23*</td>
</tr>
<tr>
<td>2003</td>
<td>247</td>
<td>271</td>
<td>24*</td>
</tr>
<tr>
<td>2005</td>
<td>247</td>
<td>270</td>
<td>23*</td>
</tr>
</tbody>
</table>


*p < .05*
Gender

The National Council for Educational Statistics’ National Assessment of Educational Progress (NAEP, 2005) has conducted nationally representative standardized assessments of reading achievement for U.S. students in grades 4 and 8 for over 30 years. Data from the NAEP have indicated a significant achievement gap between male and female students in every state in the nation in grades 4 and 8 since eligibility data have been collected. Females have consistently outperformed males across grade levels in reading (see Table 6).

These data from NAEP are supported by numerous studies that have demonstrated a reading achievement gap in which girls have outperformed boys. For example a comprehensive study of gender differences in reading comprehension identified higher scores for girls across the varied cultures represented in 43 countries (Chiu & McBride-Chang, 2006). Epidemiological studies of the incidence of dyslexia and other reading problems have repeatedly shown that boys are disproportionately identified with these developmental problems (e.g., James, 1992; Muter, 2003; Stein, 1994).
Table 6

Grades 4 and 8 Mean Reading Scaled Scores by Gender for Available Years 1998-2005

<table>
<thead>
<tr>
<th>Year</th>
<th>scaled score males</th>
<th>scaled score females</th>
<th>achievement gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>217</td>
<td>212</td>
<td>5*</td>
</tr>
<tr>
<td>2000</td>
<td>219</td>
<td>208</td>
<td>11*</td>
</tr>
<tr>
<td>2002</td>
<td>222</td>
<td>215</td>
<td>7*</td>
</tr>
<tr>
<td>2003</td>
<td>222</td>
<td>215</td>
<td>7*</td>
</tr>
<tr>
<td>2005</td>
<td>222</td>
<td>216</td>
<td>6*</td>
</tr>
<tr>
<td>Grade 8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>270</td>
<td>256</td>
<td>14*</td>
</tr>
<tr>
<td>2002</td>
<td>269</td>
<td>260</td>
<td>9*</td>
</tr>
<tr>
<td>2003</td>
<td>269</td>
<td>258</td>
<td>11*</td>
</tr>
<tr>
<td>2005</td>
<td>267</td>
<td>257</td>
<td>10*</td>
</tr>
</tbody>
</table>


*p < .05
Disability

The No Child Left Behind Act (NCLB) of 2001 requires states and school districts to set achievement targets for all students, called adequate yearly progress (AYP). Determinations of AYP are required for all students as an aggregate and also for key subgroups of students identified by race, economic disadvantage, disability, and English language learner status (U.S. Department of Education Office of Elementary and Secondary Education, 2002). Substantial achievement gaps between students with disabilities and their nondisabled counterparts in U.S. schools has raised the awareness that students with disabilities also experience higher rates of unemployment and incarceration later in adulthood (U.S. Department of Education Institute of Education Sciences, 2006). These achievement gaps exist in both reading and mathematics.

Studies of the reading achievement levels among students with disabilities and the general population of students in Ohio schools indicate that a significant achievement gap exists between these populations (State Board of Education Closing Achievement Gaps Task Force, 2003). In addition, a disproportionate number of students with disabilities drop out of Ohio high schools compared to their nondisabled peers.

Oral Reading Fluency

Information processing theories of cognitive development gave rise to theoretical hypotheses about the differences in reading comprehension skills among good and poor readers. LaBerge and Samuels (1974) developed the hypothesis that automaticity of reading decoding processes allowed for higher levels of reading comprehension because the cognitive system could devote more attention resources to comprehension when word
recognition was effortless. This conceptualization of automaticity in word recognition as the foundation of good reading comprehension became the basis for investigating the predictive validity of oral reading fluency in relation to comprehension.

Systematic measures of oral reading fluency were introduced as part of the curriculum based measurement (CBM) movement (Deno, 1985; Fuchs & Fuchs, 2002; Shinn, 1989, 2002) which measured academic proficiency using basic foundational skill targets. Measuring oral reading fluency involves listening to the student read a selected passage from grade-level reading material for a one minute interval and counting the number of words read correctly. Oral reading fluency has been studied extensively as a curriculum-based measure and has the greatest empirical evidence in the CBM literature (Baker, et al., 2008). Curriculum-based measurement techniques such as oral reading fluency are efficient, simple, standard assessments of basic skills that are designed to monitor the progress of students over time in relation to the progress of typical students in key curriculum areas (Shinn, 1998). Because they are short-duration assessments, CBM procedures are vital tools used for monitoring student progress through phases of intervention (Chidsey-Brown & Steege, 2005).

Reading comprehension can also be measured using CBM procedures. Researchers have utilized several techniques for measuring reading comprehension such as recall of information, answering questions, vocabulary identification, and cloze tests (i.e., identification of missing words in passages). Although these measures are available, validity studies of oral reading fluency have reported correlations with published measures of reading comprehension (e.g., IOWA Test of Basic Skills and state mandated
reading assessments) in the range of .71 to .91 among students in the primary grades (Deno, Mirkin & Chiang, 1982). These relationships were stronger than other accepted components of reading such as vocabulary identification or cloze procedures. Fuchs, Fuchs and Maxwell (1988) studied the predictive validity of oral reading fluency procedures among middle school students receiving special education services. They reported that oral reading fluency correlated more strongly with reading comprehension (.92) than with reading decoding (.81). This finding that oral reading fluency is more strongly related to reading comprehension than to reading decoding was also reported by Shinn, Good, Knutson Tilly and Collins (1992) and Shinn (1998). Because of these findings and the simplicity of administration, measures of oral reading fluency are a strong index of overall reading proficiency (Baker et al., 2008).

The No Child Left Behind Act (2001) requires states and school districts to determine if adequate yearly progress is being made by students beginning in grade 3 utilizing state proficiency testing measures in reading and mathematics. Validity studies of oral reading fluency measures as predictors of state proficiency test results have yielded moderate to strong correlations. For example, Shaw and Shaw (2002) reported correlations ranging from .73 to .80 between oral reading fluency measures and the grade 3 Colorado Student Assessment Program reading test. Wood (2006) reported correlations of .67 to .75 with ORF measures and the grade 4 and 5 Colorado proficiency tests. Correlations of .49 to .83 were reported between oral reading fluency measures and the reading portion of the Michigan Educational Assessment Program (McGlinchey & Hixson, 2004). Similar results have been reported in validity studies with the North
Carolina and Arizona reading tests (Barger, 2003; Wilson, 2005), the Ohio Reading Proficiency Tests (Vander Meer, Lentz, & Stollar, 2005), and the Oregon Statewide Reading Assessment (Baker et al., 2008).

Summary

The current study proposes to address the presence of confounding covariates such as race, economic disadvantage, prior achievement, gender, and oral reading fluency skills in a local program evaluation of a standard-protocol Tier-2 reading intervention program. School psychologists working in the role of systems consultants or intervention program evaluators (Ysseldyke et al., 2006) have traditionally utilized adjusted means approaches such as Analysis of Covariance (ANCOVA) or multiple regression techniques to determine the effectiveness of interventions when known confounding covariates have been demonstrated (e.g., Aarnoutse & Schellings, 2003; Balfanz, Legters, & Jordan, 2004; Gettinger and Stoiber, 2007; Modaressi & Wolanin, 2007; Nokes, Dole, & Hacker, 2007; Vaughn et al., 2009; Wheldall, 2000). As this review of literature has described, there are significant weaknesses in the use of adjusted means approaches, such as undefined validity of the adjusted achievement construct (Newman et al., 2005). The current study proposes to conduct a program evaluation of a standard-protocol Tier-2 reading intervention using the dual analyses of ANCOVA and an alternate technique, propensity score analysis (PSA) (D’Agostino, 1998; Rosenbaum, 2002; Rosenbaum & Rubin, 1983, 1984). The PSA technique makes the adjustment for confounding covariates on the independent variable instead of the dependent variable, and eliminates the need for a construct of adjusted achievement (Tracz, Nelson, Newman, & Beltran,
2005). This study will compare the results of these dual analyses and determine if the analyses result in similar or different conclusions regarding the effectiveness of the intervention.
CHAPTER III

METHODOLOGY

Introduction

This study incorporated a quasi-experimental design to investigate the effectiveness of a standard protocol reading intervention program in an urban/suburban public school system (Office of Assessment, Ohio Department of Education, 2009b) in the Midwest. Specifically, it was classified as an observational study because treatment groups were assigned by a system of eligibility over which the researcher did not have control, resulting in sample bias. Students in grades three and five were assigned to traditional reading instruction or traditional instruction plus a standard protocol reading intervention. Students were selected for the reading intervention within a response-to-intervention framework incorporating a three-tiered model of intervention intensity. Within the three tiered model, all students were initially assigned to the universal intervention condition, the general curriculum. It is expected that approximately 80% of students will become proficient in reading while benefiting from this level of intervention. A second level of targeted intervention for approximately 15-20% of students who do not meet achievement benchmarks and who require a more intensive level of intervention will comprise the standard protocol reading intervention group. A third level of intervention intensity is reserved for approximately 5% of students who require the most individualized and tailored interventions. The current study concerned students in Tier 1 and Tier 2 levels of interventions.
A single dependent variable was hypothesized to be impacted by each of the two treatment conditions, four interval data covariates and two categorical covariates. The single independent variable was a measure of reading achievement gain, operationalized as the difference in scaled score points between two administrations of the Ohio Achievement Test in reading. The independent variable was participation in the Tier 1 universal instruction only, or the Tier 2 standard protocol reading intervention in addition to or participation in Tier 1 universal instruction. The first interval covariate was a measure of oral reading fluency collected prior to assignment to intervention. The second interval covariate was the baseline reading achievement level, or scaled score on the first administration of the Ohio Achievement Test in reading. The third covariate was the year of participation on intervention (2005-2006, 2006-2007 or 2007-2008). The fourth interval covariate was a nationally standardized measure of reading achievement, the IOWA Test of Basic Skills in reading. The two categorical covariates were race and economic disadvantage of participants.

A traditional method of program evaluation utilizing Analysis of Covariance (ANCOVA) was initially employed to determine if significantly greater reading gains were made by students participating in the standard protocol reading intervention than those participating only in the universal intervention. The results of this analysis permitted a comparison of the adjusted means for each group. The data were then analyzed employing Propensity Score Analysis as an alternative method of controlling for the inherent selection biases introduced by the eligibility criteria for the intervention program. Finally, the effect sizes produced by the two different methods of analysis were
compared to determine if there was a significant difference in the results of the two methods of analysis.

Participants

The participants in this study included all students in grades three and five who participated in either the Tier 1 intervention or the Tier 2 intervention during the 2005-2006, 2006-2007, and 2007-2008 school years. Because the purpose of the study was to investigate the effectiveness of Tier 2 interventions, students who participated in the Tier 3 level of intervention did not participate in the study. Overall, 573 students in third grade were included in the study. There were varying numbers of students in grade 3 for the three school years. In 2005-2006 there were 200 third grade students, in 2006-2007 there were 174 third grade students; and in 2007-2008 there were 199 third grade students. Because a grade-level building configuration exists in the school district, all of the district’s third grade students attended the same building.

Overall, 579 students in fifth grade were included in the study. There were varying numbers of students in grade 5 for the three school years. In 2005-2006 there were 198 fifth grade students, in 2006-2007 there were 180 fifth grade students; and in 2007-2008 there were 201 fifth grade students. Because a grade-level building configuration exists in the school district, all of the district’s fifth grade students also attended the same building. All of the students in the study attended the same elementary school during the intervention period investigated.

Students were included in the study if they were eligible to participate in either Tier 1 or Tier 2 interventions and if they were enrolled in the school during the period of
time data was collected for all criterion variable measures and all independent variable measures. Scores on the oral reading fluency measure were used to determine eligibility for the Tier 2 standard protocol intervention at both grade three and grade five. At grade three, students participated if they were in attendance for administration of the fall oral reading fluency measure, the fall administration of the Ohio Achievement Test in reading, the Iowa Test of Basic Skills reading test, and the spring administration of the Ohio Achievement Test in reading. At grade five, students participated if they were in attendance for administration of the fall oral reading fluency measure, the prior year (fourth grade) administration of the Ohio Achievement Test in reading, the Iowa Test of Basic Skills reading test, and the spring administration of the Ohio Achievement Test in reading.

Of the 573 third grade students who were enrolled in the school during the school years 2005-2006, 2006-2007, and 2007-2008, there were 166 who were eligible for the Tier 2 reading intervention. In 2005-2006, there were 59 eligible students; in 2006-2007, there were 48 eligible students; and in 2007-2008, there were 59 eligible students.

Of the 579 fifth grade students who were enrolled in the school during the school years 2005-2006, 2006-2007, and 2007-2008, there were 136 who were eligible for the Tier 2 reading intervention. In 2005-2006, there were 38 eligible students; in 2006-2007, there were 39 eligible students; and in 2007-2008, there were 59 eligible students. In each year, the targeted population eligible for the Tier 2 intervention earned the lowest 20% of oral reading fluency scores. Throughout the course of each school year,
transience affected the exact percentage of students who were enrolled in the Tier 2 intervention at both grade levels.

**Exclusion of Potential Participants**

There were a number of criteria that caused potential participants to be excluded from the study. At grade three, students were excluded if they were not enrolled during collection of both fall and spring administrations of the Ohio Achievement Test in reading. Students in grade three were also excluded if they were not present for one of the covariate measures of oral reading fluency or the IOWA Test of Basic Skills in reading. Because there was a concerted effort to administer these measures on a make-up basis for students absent during the regularly scheduled administrations, the students excluded were almost exclusively those who transferred into or out of the district during the third grade year.

At grade five, students were excluded if they were not enrolled during the spring administration of the Fourth Grade Ohio Achievement Test in reading, because it was used as the baseline measure for determining gain scores. They were also excluded if they were not present for the spring administration of the Fifth Grade Ohio Achievement Test in reading or one of the covariate measures of oral reading fluency or the IOWA Test of Basic Skills in reading. Similar to the third grade population, there was a concerted effort to administer these measures on a make-up basis for students absent during the regularly scheduled administrations. The students excluded were almost exclusively those who transferred into or out of the district during the fourth or fifth grade year of school. The number of students participating in the study, the number
excluded due to transience, and students’ treatment group eligibility are summarized in Table 7.

Table 7

*Participants Before and After Exclusion Procedures*

<table>
<thead>
<tr>
<th>Year</th>
<th>Total Enrollment</th>
<th>After Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tier 1</td>
<td>Tier 2</td>
</tr>
<tr>
<td>Grade 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005-2006</td>
<td>200</td>
<td>59</td>
</tr>
<tr>
<td>2006-2007</td>
<td>174</td>
<td>48</td>
</tr>
<tr>
<td>2007-2008</td>
<td>199</td>
<td>59</td>
</tr>
<tr>
<td>Totals</td>
<td>573</td>
<td>166</td>
</tr>
<tr>
<td>Grade 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005-2006</td>
<td>198</td>
<td>38</td>
</tr>
<tr>
<td>2006-2007</td>
<td>180</td>
<td>39</td>
</tr>
<tr>
<td>2007-2008</td>
<td>201</td>
<td>59</td>
</tr>
<tr>
<td>Totals</td>
<td>579</td>
<td>136</td>
</tr>
</tbody>
</table>
Instrumentation

Three published measures of reading achievement were utilized in the data collection procedures for the study, Dynamic Indicators of Basic Early Literacy, the IOWA Test of Basic Skills Reading Test, and the Ohio Achievement Test in Reading.

DIBELS

The Dynamic Indicators of Basic Early Literacy (DIBELS; Kaminski & Good, 1996) is a commonly utilized, curriculum-based benchmark assessment in reading, one of many tools available to school psychologists for progress monitoring (Burns & Coolong-Chaffin, 2006; Brown-Chidsey & Steege, 2005). The DIBELS reading assessments measure prereading skills at grades K-1 and oral reading fluency skills at grades 1-6 (Brown-Chidsey & Steege, 2005). The DIBELS assessments include measures of initial sound fluency (ability to recognize and produce the initial sound in an orally presented word) at grade K; phoneme segmentation fluency (ability to segment three- and four-phoneme words into their individual phonemes) at grades K-1; nonsense word fluency (ability to orally read non-words) at grades K-2; and oral reading fluency in grades 1-6 (University of Oregon Center on Teaching and Learning, 2009; Jenkins, Hudson, & Johnson, 2007).

Each student’s classroom teacher measured students’ oral reading fluency by listening to the student read a selected passage from grade-level reading material for a one minute interval and counting the number of words read correctly. Because the progress of all students was monitored at least three times per year (fall, winter, spring), any student who demonstrated a poor response to universal intervention was identified
for inclusion in Tier 2 intervention activities. Poor response to the universal intervention is typically operationalized as failure to meet the benchmark criteria at grade level identified by the curriculum-based measure (Jenkins et al., 2007; Burns & Coolong-Chaffin, 2006). The benchmark criteria are based on both national normative criteria and local norms for the average performance of students in the same grade at the child’s school (Brown-Chidsey & Steege, 2005). In the current study, students with the lowest 20% of oral reading fluency scores for the grade level were assigned to Tier 2 intervention.

*IOWA Test of Basic Skills – Reading*

The IOWA Test of Basic Skills (ITBS) (Form B) – Reading (Riverside Publishing, 2007) is a nationally standardized test of reading skills that is part of a complete battery of academic assessments designed for group administration. Reading scores are derived from student performance on three subtests at the third and fifth grade levels: vocabulary, word attack and comprehension. Each of the subtests is timed, and no subtest is longer than 30 minutes. At the third and fifth grade levels, each test contains reading passages of varying length and difficulty. At each test level, there is at least one narrative, a poem, and one passage about a science or social studies topic.

Scores on the ITBS Reading are reported as scaled scores and derived scores such as normal curve equivalents, percentiles or developmental age scores. The publisher reports internal consistency reliability estimates of 0.94 for the complete battery. A band of 12 percentile rank points is reported as a confidence band for the reading assessment at a 95% confidence level (Riverside Publishing, 2007).
Ohio Achievement Test – Reading

The third and fifth grade Ohio Achievement Test in reading are standards-based assessments of reading skills based on the Ohio Content Standards. The test format includes 36 or 37 test items assessing student skills in the standards of acquisition of vocabulary, reading applications for both literary and informational text, and reading processes. Students are presented with four reading selections of approximately 350-500 words and assessed in the four content areas using multiple choice (80% of items), written short answer (14% of items), and written extended response items (6% of items). Test results are reported in scaled scores and raw scores in each of the content areas (Office of Assessment, Ohio Department of Education, 2005).

The most current technical data reported by ODE for the fifth grade test indicated a mean scaled score of 413.78 and standard error of measurement of 10.73 scaled score points. A reliability coefficient of 0.88 was reported, but it was not clear whether this was a measure of internal consistency reliability, test-retest reliability or alternate forms reliability (Office of Assessment, Ohio Department of Education, 2009a). Technical data for the third grade test indicated a mean scaled score of 414.04 and a standard error of measurement of 9.67 scaled score points. A reliability coefficient of 0.86 was reported for the third grade test, although the type of reliability measure was unspecified (Office of Assessment, Ohio Department of Education, 2009a).
Procedure

Assignment of Participants

A response-to-intervention framework was utilized to assign students to treatment groups and deliver the intervention under investigation. This delivery system has a long history in the public health and prevention literature (Ysseldyke et al., 2006). Within this framework, interventions are delivered to students in three levels: universal, targeted, and intensive to accomplish the goal of closing achievement gaps and meeting adequate yearly progress goals (Marston et al., 2003). The first tier, universal intervention, is designed to meet the needs of approximately 85% of learners, and involves a challenging research-based general curriculum. The second tier, targeted, represents interventions for 10-20% of students who have not met achievement benchmarks and who require a more intensive level of intervention. The third tier, intensive, represents approximately 5% of students who have not adequately responded to Tier 2 targeted interventions, and who therefore require highly individualized and tailored interventions, including students with disabilities (Burns et al., 2007).

All students in the elementary school who participated in the general curriculum were initially assigned to Tier 1 (universal) intervention. Although all students continued in this level of intervention for the duration of the study, a subset of students was assigned also participate in Tier 2 intervention, an additional level of supplemental instruction. Students were assigned to the Tier 2 level of (targeted) intervention as a result of ongoing universal screenings of reading fluency. Universal screening occurred three times annually (i.e., October, January, and May) for all students using the DIBELS
oral reading fluency measure. Students were assigned to Tier 2 intervention if their oral
reading fluency score was in the lowest 20% of their grade level class. Scores from the
May DIBELS administration were used to identify the lowest 20% of the grade level
class for fall intervention. Students who enrolled in the school after the May
administration were screened during their first week of attendance and their scores were
included in the eligibility determination. In this way, it was possible to assign students to
Tier 2 intervention after the first week of the school year.

A strategy for responding to student transience was incorporated in the study in
order to provide targeted intervention to as many students as possible on an ongoing
basis. As new students enrolled in the school, each was screened using the DIBELS oral
reading fluency measure. As students in the Tier 2 intervention group withdrew from the
school as part of the transfer process, new student enrollments whose DIBELS scores
were within the lowest 20% of the class were assigned to Tier 2 intervention. Students
were classified as receiving the Tier 2 intervention if they participated for at least 27
weeks.

*Intervention Procedures*

Instruction of all students was conducted by teachers licensed appropriately for
the corresponding grade level assignment. Universal intervention was provided by each
child’s classroom teacher in the student’s assigned general education classroom. All
students benefited from delivery of the same universal curriculum. Students assigned to
Tier 2 intervention received supplemental instruction provided by teachers licensed
appropriately for the corresponding grade level assignment. Tier 2 intervention was
delivered primarily in small tutoring rooms near the students’ general education classrooms, although some instruction was provided to this group of students within the general education classroom in a small group format. Students assigned to Tier 2 intervention received supplemental instruction four to five times per week for 30-40 minutes in small groups of three to five students, consistent with practices demonstrated in other evidence-based reading interventions in the literature (Burns et al., 2006; Burns & Coolong-Chaffin, 2006; Vaughn & Roberts, 2007; Wanzek & Vaughn, 2007). Intervention was provided each week that school was in session during the 2005-2006, 2006-2007, and 2007-2008 school years, for a total of at least 36 weeks per school year.

The instructional activities provided in the Tier 2 intervention followed the recommendations of the National Reading Panel (2000) for early reading that include explicit instruction in (a) phonemic awareness, (b) systematic phonics, (c) reading fluency, (d) vocabulary development, and (e) comprehension. In order to accomplish these, tutoring sessions included phonemic segmentation and phonemic construction, repeated oral reading, application of phonological skills to decodable text, and use of comprehension strategies including question answering, summarization and question generation.

Throughout the school year, progress monitoring of oral reading fluency was conducted using DIBELS. The progress of all students in the Tier 1 intervention was monitored three times during the school year (i.e., October, January, and May). The oral reading fluency of students in the Tier 2 intervention was monitored weekly using the same DIBELS measure.
Data Collection/Coding Procedures

For the purpose of this study, each student was assigned a student identification number in order to protect confidentiality. Data collected for each student included: (a) the type of treatment, either Tier 1 or Tier 2 intervention, (b) DIBELS scores of oral reading fluency prior to the intervention, (c) IOWA Test of Basic Skills reading composite scaled score, (d) baseline Ohio Achievement Test reading score, (e) outcome score for the Ohio Achievement Test in reading, (f) the student’s race, (g) the student’s socioeconomic status, and (h) the student’s year of participation.

Each student’s identification number and other data pertinent to this study were transferred onto a spreadsheet. This spreadsheet was then entered into a computer database, in order that statistical analyses of the data could be performed using programs found in SPSS Base 13.0 for Windows (SPSS, 2004). Data included: (a) student identification number, (b) type of intervention, (c) oral reading fluency score, (d) IOWA Test of Basic Skills reading composite scaled score (e) baseline scores on the Ohio Achievement Test in reading, (f) outcome score on the Ohio Achievement Test in reading, (g) gain score for the Ohio Achievement Test in reading, (h) the student’s race, (i) the student’s socioeconomic level, and (j) the student’s year of participation.

For the purpose of this study, gain scores were derived by computing the arithmetic difference between two administrations of the Ohio Achievement Test in reading. This was done by subtracting the most recent scaled score from the scaled score achieved on the baseline measure. For third grade students, both administrations occurred during the third grade school year. During the 2005-2006 school year, the
baseline measure was administered in October and the outcome measure was administered in March. During the 2006-2007 and 2007-2008 school years, the baseline measure was administered in October and the outcome measure was administered in May as determined by the Ohio Achievement Test Administration Schedule, Office of Assessment (2008). For fifth grade students, the baseline measure was the administration of the fourth grade Ohio Achievement Test in reading and the outcome measure was the administration of the fifth grade Ohio Achievement Test in reading. During the 2005-2006 school year, the baseline administration was administered in March, 2005 and the outcome measure was administered in March, 2006. During the 2006-2007 school year, the baseline measure was administered in March, 2006 and the outcome measure was administered in May, 2007. During the 2007-2008 school year, the baseline measure was administered in May, 2007 and the outcome measure was administered in May, 2008.

In order to increase the statistical power in determining possible differences between treatment groups, data were pooled for three cohorts of students within each grade level for the 2005-2006, 2006-2007, and 2007-2008 school years. In order to determine if any cohort effects introduced additional bias in the data, each analysis tested the inclusion of the year of participation as a potential covariate.

**Statistical Analysis**

A traditional method of program evaluation utilizing Analysis of Covariance (ANCOVA) was initially employed to determine if significantly greater reading gains were made by students participating in the standard protocol reading intervention than those participating only in the universal intervention. The results of this analysis
permitted a comparison of the adjusted means for each group. The data were then analyzed employing Propensity Score Analysis as an alternative method of controlling for the inherent selection biases introduced by the eligibility criteria for the intervention program. Finally, the effect sizes produced by the two different methods of analysis were compared to determine if there was a significant difference in the results of the two methods of analysis.

*Analysis of Covariance Methodology*

Analysis of Covariance (ANCOVA) is an analytic method commonly utilized by educational researchers when random assignment of subjects is not possible, such as in studies involving comparison of intact classrooms, or when systematic difference in experimental and control groups are evidenced by group differences on a pretest measure (Cohen & Cohen, 1983; Huitema, 1980; Pedhazur, 1997; Stevens, 2007). In this method, a number of variables known to be correlated with the dependent variable are statistically removed from the error variance term. The logic of this method is that if the lack of random assignment to treatment groups is impossible, it is wise to remove the known bias from the equation. This theoretically allows the researcher to test the amount of unique variance in the dependent variable accounted for by treatment group membership alone. The method involves first identifying the covariates that may introduce bias in testing differences based on group membership. For example, in a study evaluating the effectiveness of reading interventions, the evaluator may determine that students assigned to the intervention group are disproportionately male, economically disadvantaged, African American, and have low oral reading fluency skills. Next, the
magnitude of relationship with the dependent variable and with other potential covariates is determined through correlational analysis. The covariates for the ANCOVA analysis are then selected based upon their relative magnitude of correlation with the dependent variable and their ability to identify unique quantities of the variance relative to the other covariates. These variances are then statistically removed to reduce systematic bias from nonrandom group assignment and to reduce the overall within group or error variance (Stevens, 2007). The result is a comparison of group means that have been adjusted to “what they would be if all groups started out equally on the covariates” (Stevens, 2007, p. 309).

**Propensity Score Analysis Methodology**

Rosenbaum (2002) and Rosenbaum and Rubin (1983, 1984) developed a method of analyzing the differences between the means of non-randomized groups using propensity scores. It is a way of adjusting the comparison of groups when there is systematic bias between treatment groups on a set of covariates (Tracz et al., 2005). Whereas adjusted means approaches such as multiple regression and ANCOVA try to adjust for bias by making an adjustment on the dependent variable, propensity score analysis attempts to adjust for this bias by making the adjustment on the independent variable (Yanovitsky, Zanutto, & Hornik, 2005). The technique can be used to make causal inferences from observational studies and to determine if a dataset can be utilized to address causal questions (Rubin, 1997; Perkins et al., 2000). It is also useful in eliminating bias that is inherent when comparing treatment groups in observational
studies. Fundamentally, propensity score analysis is a measure of differences between group means (D’Agostino, 1998).

The analysis was developed to adjust for the selection bias in group assignment based on a set of covariates. The propensity score is a method of summarizing the information required to balance the distribution of all of the covariates in a single scale. From this scale of propensity scores, subjects are divided into subclasses in both the treatment and control conditions that serve to balance all of the covariates. Rosenbaum and Rubin (1984) reported that distribution of subjects into five subclasses is sufficient to remove more than 90% of the bias attributable to each of the covariates. Modeling the set of propensity scores is done before testing the effects on a dependent variable, and therefore independent of any effects of the outcome. The goal of using propensity scores is to approximate random assignment by balancing the subjects in treatment and control groups on the set of covariates (Yanovitzky et al., 2005).

The propensity score technique allows school psychologists to address the issue of selection bias due to confounding covariates when evaluating the effectiveness of Tier 2 interventions. Most importantly for the purposes of this study, because this is done independent of the outcome variable, there is no need to make statistical adjustments on the outcome variable means. As a result, the construct of reading achievement remains unchanged. This allows the school psychologist in the role of program evaluator to more effectively match the analytic tool and the research question. The school psychologist avoids committing a Type VI error in which the research question addresses reading achievement and the analytic technique addresses the analysis of adjusted reading
The propensity score is a conditional probability that an individual subject will be assigned to a treatment group given a known set of confounding variables or covariates (Rosenbaum & Rubin, 1984). In most research using a quasi-experimental research design, subjects in treatment and control groups may differ in the observed values of a set of covariates and in the number of covariates affecting subjects. If a regression-type approach is used to control for these confounding variables, the researcher assumes the same number of covariates affect both treatment and control groups. This may lead to biased results (Yanovitzky et al., 2005). By developing a model of the scalar function of the covariates, the school psychologist can compute propensity scores that summarize each set of the covariates for all subjects before comparing experimental and control groups (Rosenbaum & Rubin, 1984). The propensity scores can then be used to assign subjects to subclassification levels in order to compare treatment and control groups. In this way, propensity score analysis is a method of compensating for lack of random assignment in quasi-experimental research (D’Agostino, 1998; Hahs-Vaughn & Onwuegbuzie, 2006; Rosenbaum, 2002).

Estimating the Propensity Score

The most common method of computing propensity scores when subjects have been assigned to two groups (i.e., treatment and control) involves utilizing the logistic regression model (Fraas, Newman, Bagakas, & Newman, 2006). Logistic regression is
appropriate for analyzing an outcome measure that is dichotomous (LeBlanc & Fitzgerald, 2000) when a set of variables are known to be related to that outcome. In this case, logistic regression is used to compute the probability that each subject would have been assigned to the treatment or control group based on a set of covariates.

*Steps Used to Conduct Propensity Score Analysis*

Propensity score analysis can be understood by reviewing the sequence of steps used to conduct the analysis. Yanovitzky et al., (2005) described six steps researchers should follow to conduct a propensity score analysis.

*Step 1: Identify and select the covariates.* The researcher must select a set of covariates based on theoretical grounds and previous empirical studies. This should be done without an analysis of the intercorrelations of variables in the data set under investigation. The covariates should be selected based upon the known biases in the population that have been documented in the literature. These covariates are used to estimate the propensity scores used to form sub-groups of participants. In the current study, there is an extensive literature demonstrating disproportionately poor reading achievement among economically disadvantaged students (e.g., Bradley & Corwyn, 2002; Brooks-Gunn et al., 1999; Chiu & McBride-Chang, 2006; Ogle et al., 2003; State Board of Education Closing Achievement Gaps Task Force, 2003), African-American students (e.g., State Board of Education Closing Achievement Gaps Task Force, 2003; U.S. Department of Education Institute of Education Sciences, 2005), and students with poor reading fluency (e.g., Deno, Mirkin & Chiang, 1982; Fuchs, Fuchs, & Maxwell, 1988; Shinn, 1998; Shinn, Good, Knutson Tilly & Collins 1992).
Step 2: Assess the initial imbalance in the covariates. The school psychologist determines if there is an initial imbalance in each of the covariates based upon group assignment. For covariates reported in an interval level of measurement, an independent-samples $t$ test can be used. Imbalances on covariates reported as dichotomous variables can be analyzed using a chi square test of differences in proportions. Assessing the initial imbalance in the covariates allows the school psychologist to determine if the balance is similar to the amount of balance one would expect in a completely randomized experiment (Rosenbaum & Rubin, 1984; Zanutto, Lu, & Hornik, 2005). If the balance is adequate, it is not necessary to apply the propensity score analysis technique, and the group means on the outcome variable can be analyzed directly. Determining the amount of initial imbalance on the covariates also provides a benchmark for measuring any increase in balance gained by utilizing the propensity score methodology (Yanovitzky et al., 2005).

Step 3: Calculate the propensity scores. If an imbalance exists between treatment and control groups on one or more of the covariates, the school psychologist should estimate the propensity scores for each student in the study. These propensity scores can be estimated using a variety of methods. The most common method of computing propensity scores is to use a logistic regression model, with the dependent variable being the dichotomous outcome variable (i.e., experimental and control groups) and the covariates identified as the independent variables (D'Agostino, 1998; LeBlanc & Fitzgerald, 2000; Rosenbaum & Rubin, 1984).
Step 4: **Stratify the propensity scores into four or five subclassifications.** Once the propensity scores are estimated, they are stratified into four or five levels with equal or nearly equal numbers of subjects in the categories. The researcher is encouraged to utilize judgment in assigning cases to strata. For example, it is important to include all cases with the same propensity score in the same stratum, and it is desirable to divide strata between cases that have dissimilar propensity scores (D’Agostino, 1998). The number of subjects in treatment and control groups will typically be unequal in all strata, with greater disparity in numbers of subjects in the highest and lowest strata. As reported by Cochran (1983), there is virtually no benefit to stratifying the propensity scores in more than four or five levels.

Step 5: **Assess the balance on the covariates across the treatment groups.** After the propensity scores have been stratified into four or five subclassifications, the school psychologist can determine the degree to which the propensity score groups have removed initial bias on the covariates. This can be done by using a two-way analysis of variance (ANOVA), where the two factors are the treatment groups (i.e., Tier 1 and Tier 2) and the propensity score groups (i.e., subclassifications 1 through 5) and each covariate is used as a criterion variable (Yanovitzky et al., 2005). The participants have been balanced on the covariates when neither the treatment main effect nor the interaction effect is statistically significant.

Step 6: **Estimate and statistically test the difference between the treatment means.** The overall treatment effect is determined in a two step process. First, the difference between the treatment means on the outcome variable (i.e., reading achievement) are
statistically tested for each propensity score subclassification using independent $t$ tests for unequal groups. Second, the overall treatment effect computed across the five propensity score subclassifications. Yanovitzky et al. (2005) suggested that the overall estimate of the treatment effect can be calculated by averaging the differences between means of the treatment groups across all propensity score groups. The overall treatment effect is calculated using Equation 1 as follows:

$$\hat{\delta} = \sum_{k=1}^{4} \frac{n_k}{N} (\overline{Y}_{ek} - \overline{Y}_{ck})$$

Where:

1. The estimated overall treatment effect is $\hat{\delta}$.
2. The propensity score subclassifications (4 or 5) are represented by $k$.
3. The total number of participants is $N$.
4. The number of participants in a propensity score subclassification $k$ is $n_k$.
5. The mean of the criterion variable for the experimental group is $\overline{Y}_{ek}$ and the mean of the criterion variable for the control group is $\overline{Y}_{ck}$ within each propensity score subclassification.

The estimated standard error for the estimated treatment effect is calculated using Equation 2 as follows:

$$\hat{s}\left(\hat{\delta}\right) = \sqrt{\frac{\sum_{k=1}^{4} n_k^2}{N^2} \left( \frac{s_{ek}^2}{n_{ek}} + \frac{s_{ck}^2}{n_{ck}} \right)}$$
Where:

1. The number of participants in the $k$ propensity score subclassification is $n^2_k$.
2. The total number of participants is $N$.
3. The sample variances of the experimental and control groups are $s^2_{ek}$ and $s^2_{ck}$, respectively, in subclassification $k$.
4. The number of participants in the experimental group is $n_{ek}$ and the number of participants in the control group is $n_{ck}$.

The $t$ test value for the estimated treatment effect is calculated by dividing the estimated treatment effect ($\hat{\delta}$) by the standard error value($\hat{s}(\hat{\delta})$).

**Comparing Treatment Effects Determined by ANCOVA and Propensity Score Analysis Methodologies**

Comparing the treatment effect sizes determined by Analysis of Covariance (ANCOVA) and propensity score analysis methodologies first required conversion of test statistics for each of the program evaluations to a common standardized measure of effect size. Second, these effect sizes were statistically compared to determine if there was a significant difference between the effect sizes reported by each analytic technique.

The standardized effect size calculation was computed using the formula for Hedges’ $g$ (Hedges, 1981) using the calculated $t$ values and sample sizes from the ANCOVA and propensity score analyses. The following formula was used to compute the standardized effect size values:
The resulting standardized effect sizes were classified according to numerical criteria suggested by Cohen (1988). Standardized effect sizes of .20 to .49 are classified as small, .50 to .79 as medium, and .80 and above as large. A general comparison of the effect sizes was conducted by a simple qualitative comparison based on these classification criteria.

Statistical comparison of the two standardized effect sizes was conducted using the procedure described by Rosenthal (1991). Two standardized effect sizes were compared using a Fisher’s Z test when the estimated variance of Hedges’ $g$ is used as a measure of within-groups variance term in the denominator. The result is a single $Z$ statistic that describes the probability that the effect sizes belong to the same sampling distribution of effect sizes. This estimate of the variance of $g$ is computed using the following formula:

$$w = \frac{2(n_1n_2)(n_1 + n_2 - 2)}{(n_1 + n_2) [I^2 + 2(n_1 + n_2 - 2)]}$$

A test of the significance of the difference between the two independent $g$’s can be tested using the following formula:

$$Z = \sqrt{\frac{1}{w_A} + \frac{1}{w_B}}$$
Estimates of the variance of Hedges’ g for the ANCOVA analysis and propensity score analysis were computed using the values of t and the corresponding group sample sizes for the respective analyses. The resulting Z value tested the significance of difference between the two independent g’s.

Hypotheses for the Proposed Study

The current study proposed to address this gap in the literature by conducting a local program evaluation of a Tier 2 reading intervention program in which the unit of study is at the student level. Propensity score stratification, which utilizes more of the available data than propensity score matching, and can better account for the differences in group sizes between treatment and control groups was utilized to balance subjects on known covariates. The research questions were addressed by treating the data both by Analysis of Covariance (ANCOVA) and propensity score stratification, and a comparison of the research conclusions will be made. The value of the study was to illustrate whether there is ample evidence that school psychologists should utilize propensity score analysis rather than traditional ANCOVA to conduct local program evaluations of Tier 2 intervention programs.

Testing the Research Questions

The following research questions were addressed in this study:

1. Will third grade students who received the Tier 2 standard protocol reading intervention make different gains in reading achievement than students who received only the Tier 1 reading intervention when considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race?
This question was answered by initially computing the gain in Ohio Achievement Test reading scores for each third grade student who participated in the study. This was done by computing the difference between each student’s Ohio Achievement Test reading scores between October and May of their third grade year. Next, the mean gain in Ohio Achievement Test reading scores was computed for all third grade students. The question of difference was answered by two methods. First, an Analysis of Covariance (ANCOVA) was used to determine if third grade students receiving the Tier 2 reading intervention made different gains than students receiving only Tier 1 reading interventions when statistically controlling for the intervening variables of prior reading achievement, oral reading fluency, economic disadvantage, and race. Second, a propensity score analysis was used to determine if third grade students receiving the Tier 2 reading intervention made different gains than students receiving only Tier 1 reading interventions when controlling for prior reading achievement, oral reading fluency, economic disadvantage, and race through a process of balancing treatment groups on propensity scores.

2. Will fifth grade students who received the Tier 2 standard protocol reading intervention make different gains in reading achievement than students who received only the Tier 1 reading intervention when considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race?

This question was answered by initially by computing the gain in Ohio Achievement Test reading scores for each fifth grade student who participated in the study. This was done by computing the difference between each student’s Ohio
Achievement Test reading scores between administrations of the test in May of students’ fourth grade year and May their fifth grade year. Next, the mean gain in Ohio Achievement Test reading scores was computed across all fifth grade students. The question of difference was answered by two methods. First, an Analysis of Covariance (ANCOVA) was used to determine if students receiving the Tier 2 reading intervention made different gains than students receiving only Tier 1 reading interventions when statistically controlling for the intervening variables of prior reading achievement, oral reading fluency, economic disadvantage, and race. Second, a propensity score analysis was used to determine if fifth grade students receiving the Tier 2 reading intervention made different gains than students receiving only Tier 1 reading interventions when controlling for prior reading achievement, oral reading fluency, economic disadvantage, and race through a process of balancing treatment groups on propensity scores.

3. Will there be a significant difference in the effect sizes reported from a program evaluation of the effectiveness of the Tier 2 reading intervention among third grade students when the analysis is conducted using traditional Analysis of Covariance (ANCOVA) compared to propensity score analysis?

This question was answered by first converting the Analysis of Covariance (ANCOVA) and propensity score analysis test statistics for the third grade program evaluations to a common measure of effect size. Second the effect sizes were compared to determine if there are significant quantitative and qualitative differences between the effect sizes reported by each analytic technique.
4. Will there be a significant difference in the effect sizes reported from a program evaluation of the effectiveness of the Tier 2 reading intervention among fifth grade students when the analysis is conducted using traditional Analysis of Covariance (ANCOVA) compared to propensity score analysis?

This question was answered by first converting the Analysis of Covariance (ANCOVA) and propensity score analysis test statistics for the fifth grade program evaluations to a common measure of effect size. Second the effect sizes were compared to determine if there are significant quantitative and qualitative differences between the effect sizes reported by each analytic technique.

General Hypotheses

Hypotheses for Research Question 1: Third grade students who received the Tier 2 standard protocol reading intervention will make different adjusted achievement gains on the Ohio Achievement Test in reading \( p < .05 \) than students who received only the Tier 1 reading intervention after considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race. Third grade students who received the Tier 2 standard protocol reading intervention will make different achievement gains on the Ohio Achievement Test in reading \( p < .05 \) than students who received only the Tier 1 reading intervention when controlling for prior reading achievement, oral reading fluency, economic disadvantage, and race.

Hypotheses for Research Question 2: Fifth grade students who received the Tier 2 standard protocol reading intervention will make different adjusted achievement gains on the Ohio Achievement Test in reading \( p < .05 \) than students who received only the Tier
1 reading intervention after considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race. Fifth grade students who received the Tier 2 standard protocol reading intervention will make different achievement gains on the Ohio Achievement Test in reading (p<.05) than students who received only the Tier 1 reading intervention when controlling for prior reading achievement, oral reading fluency, economic disadvantage, and race.

Hypothesis for Research Question 3: There will be a significant difference (p<.05) between the effect sizes estimating the difference between Ohio Achievement Test Reading gains for third grade students who participated in the Tier 2 intervention compared to students who received only the Tier 1 reading intervention when the differences are estimated using ANCOVA and propensity score analysis.

Hypothesis for Research Question 4: There will be a significant difference (p<.05) between the effect sizes estimating the difference between Ohio Achievement Test Reading gains for fifth grade students who participated in the Tier 2 intervention compared to students who received only the Tier 1 reading intervention when the differences are estimated using Analysis of Covariance (ANCOVA) and propensity score analysis.
CHAPTER IV

RESULTS

Preliminary Data Analysis

*Demographic Data for Participants*

Of the 294 students in grade three for whom all data were available in the data set, the following descriptive demographic information was observed. Of this sample, 49% were female and 51% were male. The majority Caucasian students comprised 78.2% of the sample, while 21.8% were African-American. Among third grade students, 31% received free or reduced lunch (the criterion for economic disadvantage) while 69% did not receive free or reduced lunch. Among the group of students for whom all data were available, 36.1% received the Tier 2 intervention for at least 27 weeks, while 63.9% of third grade students received only Tier 1 intervention.

Of the 293 students in grade five for whom all data were available in the data set, the following descriptive demographic information was observed. Of this sample, 45.7% were female and 54.3% were male. The majority Caucasian students comprised 82.6% of the sample, while 17.4% were African-American. Among fifth grade students, 23.2% received free or reduced lunch (the criterion for economic disadvantage) while 76.8% did not receive free or reduced lunch. Among the group of students for whom all data were available, 32.1% received the Tier 2 intervention for at least 27 weeks, while 67.9% of fifth grade students received only Tier 1 intervention.
Descriptive Data

Descriptive data for the measures of continuous independent variables for third grade participants are presented in Table 8. The skewness and kurtosis values for each of the independent variables are within normal tolerances to assume normal distribution of the data on each of the variables.

Table 8

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF</td>
<td>86.06</td>
<td>37.48</td>
<td>.503</td>
<td>.313</td>
</tr>
<tr>
<td>IOWA-R</td>
<td>183.18</td>
<td>19.92</td>
<td>.021</td>
<td>-.389</td>
</tr>
<tr>
<td>OAT Fall</td>
<td>404.84</td>
<td>26.14</td>
<td>-.228</td>
<td>-.059</td>
</tr>
<tr>
<td>OAT Spring</td>
<td>421.96</td>
<td>26.72</td>
<td>-.173</td>
<td>.634</td>
</tr>
</tbody>
</table>

Application of the Analysis of Covariance (ANCOVA) procedure assumes that scores on the dependent variable are normally distributed. Nonparametric analysis of the dependent variable (i.e., reading achievement gain) among third grade students indicated no significant departure from normality in the Tier 1 intervention group (Kolmogorov-Smirnov $Z = .862, p = .447$) or in the Tier 2 intervention group (Kolmogorov-Smirnov $Z = .566, p = .906$). The assumption of homogeneity of variance on the dependent variable was evaluated using Levene’s test of equality of error variances. Results indicated no significant departure from normality, $F(1, 292) = .001, p = .976$. Use of the multiple regression procedure to conduct Analysis of Covariance should be limited to analyses in
which the relationship between predictor variables and the dependent variable is linear. Visual inspection of the dependent variable related to each independent variable suggests that the relationships are generally linear in each case.

Descriptive data for the measures of continuous independent variables for fifth grade participants are presented in Table 9. The skewness and kurtosis values for each of the independent variables are also within normal tolerances to assume normal distribution of the data on each of the variables.

Table 9

*Descriptive Statistics for Fifth Grade Participants*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF</td>
<td>117.48</td>
<td>35.94</td>
<td>.192</td>
<td>.510</td>
</tr>
<tr>
<td>IOWA-R</td>
<td>213.73</td>
<td>20.87</td>
<td>-.302</td>
<td>-.066</td>
</tr>
<tr>
<td>OAT 4</td>
<td>428.45</td>
<td>27.76</td>
<td>-.495</td>
<td>.532</td>
</tr>
<tr>
<td>OAT 5</td>
<td>430.44</td>
<td>29.41</td>
<td>-.265</td>
<td>.060</td>
</tr>
</tbody>
</table>

Application of the Analysis of Covariance (ANCOVA) procedure assumes that scores on the dependent variable are normally distributed. Nonparametric analysis of the dependent variable (i.e., reading achievement gain) among fifth grade students indicated no significant departure from normality in the Tier 1 intervention group (Kolmogorov-Smirnov $Z = .883, p = .416$) or in the Tier 2 intervention group (Kolmogorov-Smirnov $Z = .631, p = .821$). The assumption of homogeneity of variance on the dependent variable was evaluated using Levene’s test of equality of error variances. Results indicated no
significant departure from normality, $F(1, 291) = .268, p = .605$. Use of the multiple regression procedure to conduct Analysis of Covariance should be limited to analyses in which the relationship between predictor variables and the dependent variable is linear. Visual inspection of the dependent variable related to each independent variable suggests that the relationships are generally linear in each case.

Analysis of Data

Data analyses are presented within the context of each of the four research questions. A minimum alpha level of .05 was used for all statistical tests.

Research Question 1

1. Will third grade students who received the Tier 2 standard protocol reading intervention make different gains in reading achievement than students who received only the Tier 1 reading intervention when considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race?

This question was answered by initially computing the gain in Ohio Achievement Test reading scores for each third grade student who participated in the study. This was done by computing the difference between each student’s Ohio Achievement Test reading scores between October and May of their third grade year. Next, the mean gain in Ohio Achievement Test reading scores was computed for all third grade students. The question of difference was answered by two methods. First, an Analysis of Covariance (ANCOVA) was used to determine if third grade students receiving the Tier 2 reading intervention made different gains than students receiving only Tier 1 reading interventions when statistically controlling for the intervening variables of prior reading
achievement, oral reading fluency, economic disadvantage, and race. Second, a propensity score analysis was used to determine if third grade students receiving the Tier 2 reading intervention made different gains than students receiving only Tier 1 reading interventions when controlling for prior reading achievement, oral reading fluency, economic disadvantage, and race through a process of balancing treatment groups on propensity scores.

Research Question 1 Analyzed Using Analysis of Covariance (ANCOVA)

The Analysis of Covariance (ANCOVA) was conducted using SPSS in a multiple linear regression procedure that included seven potential independent variables: (a) treatment, (b) the fall (October) administration of the Ohio Achievement Test-reading, (c) the IOWA Test of Basic Skills-Reading, a norm-referenced measure of overall reading achievement, (d) race, (e) DIBELS, a measure of baseline oral reading fluency (ORF), (f) economic disadvantage, and (g) year of participation in the intervention. Treatment type was assigned as an entered independent variable and the remaining variables were entered into the model in a stepwise fashion with Ohio Achievement Test gain assigned as the criterion variable. Table 10 shows the intercorrelations among all predictor variables and the outcome variable of reading achievement gain.
When considering the individual contributions to the overall model, several covariates appeared to explain a unique proportion of the variability in reading achievement gains. These variables are listed in Table 11. Tier 1 or Tier 2 reading treatment ($R^2 = .027$), baseline OAT scores ($R^2 = .230$), standardized IOWA reading test scores ($R^2 = .049$), and race ($R^2 = .017$), each explained a significant proportion of the variability in overall reading achievement gains after statistically controlling for the effects of the other predictor variables. The unique proportions of the variability explained by oral reading fluency, $t = 1.836, p>.05$, economic disadvantage, $t = .989, p>.05$ and year of participation in intervention, $t = .662, p>.05$ were not significant.
The resulting MLR Model 4 indicated that a combination of four of the predictor variables (treatment, OAT fall, IOWA Reading, and race) explained a significant proportion of the variability in overall reading achievement gains, $F(4, 289) = 25.499$, $p < .001$. It appears that approximately 26% of the variability in overall achievement gains, $R^2 = .261$, was explained by this linear combination of predictors.

Collinearity statistics for the included variables indicate tolerance indices of .684 for treatment, .322 for OAT fall, .324 for IOWA, and .832 for race. Because tolerance indices close to 1.00 indicate absence of multicollinearity, the covariates OAT fall and IOWA should be included with awareness of some potential impact of multicollinearity. Variance inflation factors were 1.463 for treatment, 3.102 for OAT fall, 3.088 for IOWA, and 1.202 for race. The variance inflation factors are within acceptable tolerances to limit the impact of multicollinearity concerns on the results.

Table 11

Results of the ANCOVA Analysis of Grade 3 Reading Gain Scores Using MLR Model 4

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>-8.779</td>
<td>-3.275</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>OAT fall</td>
<td>-.681</td>
<td>-9.488</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>IOWA Reading</td>
<td>0.411</td>
<td>4.371</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Race</td>
<td>-7.190</td>
<td>-2.544</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Constant</td>
<td>222.454</td>
<td>10.251</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

The regression coefficient of -8.779 reported for the treatment variable in the MLR Model 4 estimated the difference between the adjusted means of the reading
achievement gains made by the Tier 1 and Tier 2 intervention treatment groups. Table 12 lists the means and standard deviations for reading gains made by both the Tier 1 and Tier 2 intervention groups. A comparison of actual means of the two groups indicates that the difference observed is not significant \( t = .284, p > .05 \), while the differences observed in adjusted achievement gains are statistically significant. The treatment coefficient obtained from the ANCOVA analysis indicates that the estimated adjusted reading achievement gain for the Tier 2 intervention group was 8.779 points smaller than the estimated adjusted reading achievement gain for the Tier 1 intervention group, \( t = -3.275, p < .01 \).

Table 12

*Descriptive Data for Third Grade Reading Achievement Gains*

<table>
<thead>
<tr>
<th>Group</th>
<th>M</th>
<th>SD</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Achievement Gain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tier 1</td>
<td>16.36</td>
<td>20.54</td>
<td>188</td>
</tr>
<tr>
<td>Tier 2</td>
<td>18.47</td>
<td>22.04</td>
<td>106</td>
</tr>
<tr>
<td>Adjusted Reading Achievement Gain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tier 1</td>
<td>20.29</td>
<td>19.70</td>
<td>188</td>
</tr>
<tr>
<td>Tier 2</td>
<td>11.51</td>
<td>20.78</td>
<td>106</td>
</tr>
</tbody>
</table>

*Hypothesis for Research Question 1 Answered Using Analysis of Covariance (ANCOVA)*

*Third grade students who received the Tier 2 standard protocol reading intervention will make different adjusted achievement gains on the Ohio Achievement Test in reading \( p < .05 \) than students who received only the Tier 1 reading intervention*
after considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race. This hypothesis was retained because the adjusted reading achievement gain for students receiving the Tier 2 intervention was significantly smaller than the adjusted reading achievement gain of students receiving the Tier 1 intervention after considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race. The treatment coefficient obtained from the ANCOVA analysis indicates that the estimated adjusted reading achievement gain for the Tier 2 intervention group was 8.779 points smaller than the estimated adjusted reading achievement gain for the Tier 1 intervention group, \( t = -3.275, p < .01 \).

Research Question 1 Analyzed Using Propensity Score Analysis

Step 1. Five covariates were identified \textit{a priori} from the general reading achievement literature as sources of significant sample bias in studies of reading interventions (Aarnoutse & Schellings, 2003; Balfanz, Legters, & Jordan, 2004; Gettinger and Stoiber, 2007; Modaressi & Wolanin, 2007; Vaughn et al., 2009; Wheldall, 2000; Nokes, Dole, & Hacker, 2007). These covariates were prior reading achievement as measured by the Ohio Achievement Test – Reading (OAT-R), Reading Gain, IOWA Test of Basic Skills - Reading (IOWA), DIBELS measure of oral reading fluency (ORF), economic disadvantage, and race. In addition to these five covariates, a dichotomous independent variable was created to identify the level of intervention as Tier 1 or Tier 2 for each participant. The mean and standard deviation values of the criterion variable (Reading Gain) and interval covariates, and the observed and expected frequencies of the
dichotomous variables for both the Tier 1 and Tier 2 intervention groups are listed in Table 13.

Table 13

*Descriptive Statistics for the Criterion Variable and Covariates – Grade 3*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Tier 1</th>
<th>Tier 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Reading Gain</td>
<td>16.36</td>
<td>20.54</td>
</tr>
<tr>
<td>OAT-R</td>
<td>414.08</td>
<td>22.74</td>
</tr>
<tr>
<td>IOWA-R</td>
<td>189.99</td>
<td>18.29</td>
</tr>
<tr>
<td>ORF</td>
<td>120.84</td>
<td>36.57</td>
</tr>
<tr>
<td>Observed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African Amer.</td>
<td>48</td>
<td>58.2</td>
</tr>
<tr>
<td>Econ. Disadv.</td>
<td>59</td>
<td>78.7</td>
</tr>
</tbody>
</table>

*Step 2.* The initial imbalances on the covariates between the Tier 1 intervention group and the Tier 2 intervention group were determined through the use of independent samples $t$ tests (for continuous variables) and $\chi^2$ tests of independence (for dichotomous variables). The results of these statistical tests are listed in Table 14 under the heading *Pre Propensity Group Formation.* Statistically significant imbalances were observed on all covariates between the Tier 1 intervention group and the Tier 2 intervention group before propensity groups were formed. Significant differences were observed in baseline OAT-R scores ($t = 11.629, p < .001$), IOWA reading scores ($t = 10.753, p < .001$, oral
reading fluency scores ($t = 10.201, p < .001$, Race ($\chi^2 = 5.441, p < .05$), and Economic Disadvantage ($\chi^2 = 20.399, p < .001$). This imbalance on covariates indicated that a simple test of differences (such as a $t$ test) between means would result in a biased statistic, and necessitated the next step in propensity score modeling.

Table 14

*Comparison of Differences between Tier 1 and Tier 2 Intervention Groups on Covariates*

*Before and After Propensity Group Formation-Grade 3*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pre-Propensity</th>
<th>Post-Propensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>OAT-R</td>
<td>$&lt;.01$</td>
<td>.929</td>
</tr>
<tr>
<td>IOWA-R</td>
<td>$&lt;.01$</td>
<td>.365</td>
</tr>
<tr>
<td>ORF</td>
<td>$&lt;.01$</td>
<td>.719</td>
</tr>
<tr>
<td>Race</td>
<td>$&lt;.05$</td>
<td>.896</td>
</tr>
<tr>
<td>Econ</td>
<td>$&lt;.01$</td>
<td>.581</td>
</tr>
</tbody>
</table>

*Step 3.* Propensity scores for each participant were estimated using logistic regression analysis. Using SPSS for the analysis, the treatment variable (i.e., Tier 1 or Tier 2 intervention group) was entered as the outcome variable and the covariates were entered as predictor variables in stepwise fashion. Within the stepwise procedure, the variables were entered by the forward method with the criterion for entry set at $p < .05$ for the Wald test value of all two-way interaction variable coefficients.
The results of the analysis of the logistic regression model, which included the predictor variables of OAT-R, IOWA, ORF, race, and economic disadvantage are listed in Table 15. The logistic regression model created from the combination of the five predictor variables explained a significant proportion of the variance in the dependent variable, group assignment ($\chi^2 = 134.707$, $df = 5$, $p < .001$). This combination of predictor variables explained a significant percentage of the variance in the dependent variable (Cox & Snell $R^2 = .368$, Nagelkerke $R^2 = .505$).

Table 15

*Results for the Logistic Regression Model- Grade 3*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Wald test value</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OAT-R</td>
<td>-.037</td>
<td>10.214</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>IOWA</td>
<td>-.025</td>
<td>2.753</td>
<td>.09</td>
</tr>
<tr>
<td>ORF</td>
<td>-.020</td>
<td>9.032</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Race</td>
<td>1.121</td>
<td>4.937</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Econ</td>
<td>.592</td>
<td>1.976</td>
<td>.16</td>
</tr>
<tr>
<td>Constant</td>
<td>19.551</td>
<td>31.889</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

*Step 4.* The resulting logistic regression model was used to estimate a probability for each of the 294 participants in the Tier 1 and Tier 2 intervention programs. Each probability value indicated the probability that the participant would be a member of the Tier 2 intervention group. These probability values are the propensity scores for each participant. The propensity scores were used to rank the participants and to classify them into four stratified groups of 73 or 74 students each.
Step 5. In order to verify that the propensity score groups removed the initial bias on the covariates, a two-way ANOVA analyses was conducted on each of the covariates. For these analyses, the two main effects were the two treatment groups (Tier 1 or Tier 2 intervention) and the four propensity score groups. The probability values for each of the two-way ANOVA treatment main effects is listed in Table 14 under the heading Post-Propensity Group Formation. The results indicated that stratifying participants into four subclassifications based upon propensity scores reduced the bias on each of the covariates to nonsignificant levels. Results of the two-way ANOVA treatment main effects for each of the covariates indicated that the treatment effects were not significant for OAT-R, $F(1, 286) = .008, p = .929$; IOWA-R, $F(1, 286) = .824, p = .365$; ORF, $F(1, 286) = .130, p = .719$; Race, $F(1, 286) = .017, p = .896$, and Econ, $F(1, 286) = .305, p = .581$.

Balance between Tier 1 and Tier 2 intervention groups is assumed to have been achieved when both the treatment main effects and the treatment-by-propensity group interaction effects are not statistically significant when tested for each covariate. Analysis of the treatment-by-propensity group interaction effects indicated that no significant interactions were observed for the covariates OAT-R, $F(3, 286) = .756, p = .520$; IOWA-R, $F(3, 286) = 2.070, p = .104$; ORF, $F(3, 286) = 2.254, p = .082$; Race, $F(3, 286) = 1.565, p = .198$, and Econ, $F(3, 286) = .521, p = .668$.

Step 6. The results of independent $t$ tests of differences between Tier 1 and Tier 2 intervention groups for each of the four propensity score strata are listed in Table 16. Statistically significant differences in reading achievement gains were observed in propensity groups 1, 2, and 3, with larger gains occurring among Tier 1 intervention
recipients. The overall treatment effect was calculated using Equation 1, and the corresponding standard error value was calculated using Equation 2. The overall estimated treatment effect was -10.871, which is shown in Table 16 by the difference between overall reading gain scores of 5.936 for the Tier 2 intervention group and 16.807 reading gain scores for the Tier 1 intervention group. Using Equation 2, the estimated standard error value for this estimated treatment effect was 3.581. The calculated $t$ value for the overall treatment effect was -3.036, which was calculated by dividing the estimated overall treatment effect of -10.871 by the estimated standard error of 3.581 was -3.036, $p < .05$. The propensity score analysis indicated that the overall treatment effect was statistically significant for the reading achievement gain scores. Students receiving the Tier 2 intervention made significantly smaller reading gains than did students in the Tier 1 reading intervention group.
Table 16

*Estimated Treatment Effects on Third Grade Reading Gain Scores Using Propensity Score Groups*

<table>
<thead>
<tr>
<th>Propensity Score Groups</th>
<th>Treatment</th>
<th>Group size</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>Tier 1</td>
<td>69</td>
<td>13.826</td>
<td>20.40</td>
<td>-1.952*</td>
</tr>
<tr>
<td></td>
<td>Tier 2</td>
<td>5</td>
<td>-4.600</td>
<td>20.06</td>
<td></td>
</tr>
<tr>
<td>Group 2</td>
<td>Tier 1</td>
<td>68</td>
<td>16.574</td>
<td>20.77</td>
<td>-2.814*</td>
</tr>
<tr>
<td></td>
<td>Tier 2</td>
<td>5</td>
<td>-10.200</td>
<td>16.08</td>
<td></td>
</tr>
<tr>
<td>Group 3</td>
<td>Tier 1</td>
<td>37</td>
<td>20.811</td>
<td>19.79</td>
<td>-2.289*</td>
</tr>
<tr>
<td></td>
<td>Tier 2</td>
<td>36</td>
<td>11.333</td>
<td>15.23</td>
<td></td>
</tr>
<tr>
<td>Group 4</td>
<td>Tier 1</td>
<td>14</td>
<td>16.071</td>
<td>22.176</td>
<td>1.715</td>
</tr>
<tr>
<td></td>
<td>Tier 2</td>
<td>60</td>
<td>27.067</td>
<td>21.468</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>Tier 1</td>
<td>188</td>
<td>16.807</td>
<td></td>
<td>-3.036**</td>
</tr>
<tr>
<td></td>
<td>Tier 2</td>
<td>106</td>
<td>5.936</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .05

**p < .01
Hypothesis for Research Question 1 Answered Using Propensity Score Analysis

Third grade students who received the Tier 2 standard protocol reading intervention will make different achievement gains on the Ohio Achievement Test in reading (p<.05) than students who received only the Tier 1 reading intervention when controlling for the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race. This hypothesis was retained because the reading achievement gain for students receiving the Tier 2 intervention was significantly smaller than the reading achievement gain of students receiving the Tier 1 intervention when controlling for the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race. The treatment coefficient obtained from the propensity score analysis indicates that the estimated reading achievement gain for the Tier 2 intervention group was 10.871 points smaller than the estimated reading achievement gain for the Tier 1 intervention group, t = -3.036, p<.01.

Research Question 2

Will fifth grade students who received the Tier 2 standard protocol reading intervention make different gains in reading achievement than students who received only the Tier 1 reading intervention when considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race?

This question was answered by initially by computing the gain in Ohio Achievement Test reading scores for each fifth grade student who participated in the study. This was done by computing the difference between each student’s Ohio Achievement Test reading scores between administrations of the test in May of students’
fourth grade year and May their fifth grade year. Next, the mean gain in Ohio Achievement Test reading scores was computed across all fifth grade students. The question of difference was answered by two methods. First, an Analysis of Covariance (ANCOVA) was used to determine if students receiving the Tier 2 reading intervention made different gains than students receiving only Tier 1 reading interventions when statistically controlling for the intervening variables of prior reading achievement, oral reading fluency, economic disadvantage, year of participation, and race. Second, a propensity score analysis was used to determine if fifth grade students receiving the Tier 2 reading intervention made different gains than students receiving only Tier 1 reading interventions when controlling for prior reading achievement, oral reading fluency, economic disadvantage, year of participation, and race through a process of balancing treatment groups on propensity scores.

Research Question 2 Analyzed Using Analysis of Covariance (ANCOVA)

The Analysis of Covariance (ANCOVA) was conducted using SPSS in a multiple linear regression procedure that included seven potential independent variables: (a) treatment, (b) the 4th grade May administration of the Ohio Achievement Test-reading, (c) the IOWA Test of Basic Skills-Reading, a norm-referenced measure of overall reading achievement, (d) race, (e) DIBELS, a measure of baseline oral reading fluency (ORF), (f) economic disadvantage, and (g) year of participation in the intervention. Treatment type was assigned as an entered independent variable and the remaining variables were entered into the model in a stepwise fashion with Ohio Achievement Test
gain assigned as the criterion variable. Table 17 shows the intercorrelations among all predictor variables and the outcome variable of reading achievement gain.

Table 17

*Correlation Matrix for All Variables – Grade 5*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Treatment</th>
<th>Gain</th>
<th>Year</th>
<th>ORF</th>
<th>IOWA</th>
<th>OAT-R</th>
<th>Econ</th>
<th>Race</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>.002</td>
<td>.044</td>
<td>-.368*</td>
<td>-.488*</td>
<td>-.462*</td>
<td>-.159*</td>
<td>.244*</td>
<td></td>
</tr>
<tr>
<td>Gain</td>
<td>-.257*</td>
<td>-.028*</td>
<td>-.167*</td>
<td>-.350*</td>
<td>.130*</td>
<td>-.193*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>.105*</td>
<td>-.041</td>
<td>.052</td>
<td>-.090</td>
<td>-.016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF</td>
<td>.639*</td>
<td>.578*</td>
<td>.256*</td>
<td>-.302*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOWA</td>
<td>.688*</td>
<td>.387*</td>
<td>-.439*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OAT-R</td>
<td>.330*</td>
<td>-.349*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Econ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.537*</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * *p<.05

When considering the individual contributions to the overall model, several covariates appeared to explain a unique proportion of the variability in reading achievement gains. These variables are listed in Table 18. Baseline OAT scores ($R^2 = .388$), standardized IOWA reading test scores ($R^2 = .143$), year of participation ($R^2 = .036$), race ($R^2 = .031$), and oral reading fluency ($R^2 = .007$), each explained a significant proportion of the variability in overall reading achievement gains after statistically controlling for the effects of the other predictor variables. The unique proportions of the
variability explained by Tier 1 or Tier 2 reading intervention, $t = -.407, p > .05$, and economic disadvantage, $t = .930, p > .05$ were not significant.

The resulting MLR Model 6 indicated that a combination of six of the predictor variables (treatment, OAT-R, IOWA Reading, year, race, and ORF) explained a significant proportion of the variability in overall reading achievement gains, $F (6, 283) = 49.404, p < .001$. It appears that approximately 51% of the variability in overall achievement gains, $R^2 = .512$, was explained by this linear combination of predictors.

Collinearity statistics for the included variables indicate tolerance indices of .728 for treatment, .474 for OAT-R, .381 for IOWA, .959 for year, .803 for race, and .540 for ORF. Because tolerance indices close to 1.00 indicate absence of multicollinearity, the covariate IOWA should be included with awareness of some potential impact of multicollinearity. Variance inflation factors were 1.373 for treatment, 2.110 for OAT-R, 2.624 for IOWA, 1.043 for year, 1.246 for race, and 1.851 for ORF. The variance inflation factors are within acceptable tolerances to limit the impact of multicollinearity concerns on the results.
Table 18

*Results of the ANCOVA Analysis of Grade 5 Reading Gain Scores Using MLR Model 6*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>-.997</td>
<td>-.407</td>
<td>.685</td>
</tr>
<tr>
<td>OAT-R</td>
<td>-.766</td>
<td>-15.003</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>IOWA Reading</td>
<td>0.689</td>
<td>9.107</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Year</td>
<td>-5.460</td>
<td>-4.566</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Race</td>
<td>-12.224</td>
<td>-4.234</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>ORF</td>
<td>.074</td>
<td>2.005</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Constant</td>
<td>11135.087</td>
<td>4.640</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

The regression coefficient of -.997 reported for the treatment variable in the MLR Model 6 estimated the difference between the adjusted means of the reading achievement gains made by the Tier 1 and Tier 2 intervention treatment groups. Table 19 lists the means and standard deviations for reading gains made by both the Tier 1 and Tier 2 intervention groups. A comparison of actual means of the two groups indicates that the difference observed is not significant $t = -.042, p > .05$. The difference observed in adjusted achievement gains is not statistically significant. The treatment coefficient obtained from the ANCOVA analysis indicates that the estimated adjusted reading achievement gain for the Tier 2 intervention group was .997 points smaller than the estimated adjusted reading achievement gain for the Tier 1 intervention group, $t = -.407, p > .05$. 
Table 19

*Descriptive Data for Grade 5 Reading Achievement Gains*

<table>
<thead>
<tr>
<th>Group</th>
<th>M</th>
<th>SD</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Achievement Gain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tier 1</td>
<td>1.95</td>
<td>23.46</td>
<td>199</td>
</tr>
<tr>
<td>Tier 2</td>
<td>2.07</td>
<td>23.60</td>
<td>94</td>
</tr>
<tr>
<td>Adjusted Reading Achievement Gain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tier 1</td>
<td>2.39</td>
<td>17.60</td>
<td>199</td>
</tr>
<tr>
<td>Tier 2</td>
<td>1.39</td>
<td>18.62</td>
<td>94</td>
</tr>
</tbody>
</table>

_Hypothesis for Research Question 2 Answered Using Analysis of Covariance (ANCOVA)_

*Fifth grade students who received the Tier 2 standard protocol reading intervention will make different adjusted achievement gains on the Ohio Achievement Test in reading (p<.05) than students who received only the Tier 1 reading intervention after considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, year of participation, and race.* This hypothesis was rejected because the adjusted reading achievement gain for students receiving the Tier 2 intervention was not significantly different than the adjusted reading achievement gain of students receiving the Tier 1 intervention after considering the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race. The treatment coefficient obtained from the ANCOVA analysis indicates that the estimated adjusted reading achievement gain for the Tier 2 intervention group was .997 points smaller than
the estimated adjusted reading achievement gain for the Tier 1 intervention group, $t = -0.407$, $p > .05$.

**Research Question 2 Analyzed Using Propensity Score Analysis**

**Step 1.** Five covariates were identified *a priori* from the general reading achievement literature as sources of significant sample bias in studies of reading interventions (Aarnoutse & Schellings, 2003; Balfanz, Legters, & Jordan, 2004; Gettinger and Stoiber, 2007; Modaressi & Wolanin, 2007; Vaughn et al., 2009; Wheldall, 2000; Nokes, Dole, & Hacker, 2007). These covariates were prior reading achievement as measured by the Ohio Achievement Test – Reading (OAT-R), Reading Gain, IOWA Test of Basic Skills - Reading (IOWA), DIBELS measure of oral reading fluency (ORF), economic disadvantage, and race. In addition to these five covariates, a nominal independent variable was created to identify the level of intervention as Tier 1 or Tier 2 for each participant. Because year of participation emerged as a significant predictor of reading gain in the ANCOVA analysis, this nominal variable was also tested in the propensity score analysis. The mean and standard deviation values of the criterion variable (Reading Gain) and interval covariates, and the observed and expected frequencies of the nominal variables for both the Tier 1 and Tier 2 intervention groups are listed in Table 20.
Table 20

*Descriptive Statistics for the Criterion Variable and Covariates – Grade 5*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Tier 1</th>
<th>Tier 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Reading Gain</td>
<td>1.95</td>
<td>23.46</td>
</tr>
<tr>
<td>OAT-R</td>
<td>437.25</td>
<td>24.52</td>
</tr>
<tr>
<td>IOWA-R</td>
<td>220.71</td>
<td>18.51</td>
</tr>
<tr>
<td>ORF</td>
<td>126.56</td>
<td>35.89</td>
</tr>
<tr>
<td>African Amer.</td>
<td>22</td>
<td>34.6</td>
</tr>
<tr>
<td>Econ. Disadv.</td>
<td>37</td>
<td>46.2</td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>71</td>
<td>68.6</td>
</tr>
<tr>
<td>2007</td>
<td>60</td>
<td>59.8</td>
</tr>
<tr>
<td>2008</td>
<td>68</td>
<td>70.6</td>
</tr>
</tbody>
</table>
Step 2. The initial imbalances on the covariates between the Tier 1 intervention group and the Tier 2 intervention group were determined through the use of independent samples $t$ tests (for continuous variables) and $\chi^2$ tests of independence (for dichotomous variables). The results of these statistical tests are listed in Table 21 under the heading *Pre Propensity Group Formation*. Statistically significant imbalances were observed on five of the covariates between the Tier 1 intervention group and the Tier 2 intervention group before propensity groups were formed. Significant differences were observed in baseline OAT-R scores ($t = 8.882, p < .001$), IOWA reading scores ($t = 9.527, p < .001$), oral reading fluency scores ($t = 6.724, p < .001$), Race ($\chi^2 = 17.402, p < .001$), and Economic Disadvantage ($\chi^2 = 7.413, p < .01$). No significant imbalance was observed based upon the year of participation in intervention ($\chi^2 = .571, p = .751$). The observed imbalance on five covariates indicated that a simple test of differences (such as a $t$ test) between means would result in a biased statistic, and necessitated the next step in propensity score modeling.
Table 21

**Comparison of Differences between Tier 1 and Tier 2 Intervention Groups on Covariates**

*Before and After Propensity Group Formation - Grade 5*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pre-Propensity Group Formation</th>
<th>Post-Propensity Group Formation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>p</em></td>
<td><em>p</em></td>
</tr>
<tr>
<td>OAT-R</td>
<td>&lt;.01</td>
<td>.183</td>
</tr>
<tr>
<td>IOWA-R</td>
<td>&lt;.01</td>
<td>.309</td>
</tr>
<tr>
<td>ORF</td>
<td>&lt;.01</td>
<td>.574</td>
</tr>
<tr>
<td>Race</td>
<td>&lt;.01</td>
<td>.731</td>
</tr>
<tr>
<td>Econ</td>
<td>&lt;.01</td>
<td>.678</td>
</tr>
<tr>
<td>Year</td>
<td>.751</td>
<td>.723</td>
</tr>
</tbody>
</table>

**Step 3.** Propensity scores for each participant were estimated using logistic regression analysis. Using SPSS for the analysis, the treatment variable (i.e., Tier 1 or Tier 2 intervention group) was entered as the outcome variable and the covariates were entered as predictor variables in stepwise fashion. Within the stepwise procedure, the variables were entered by the forward method with the criterion for entry set at *p* < .05 for the Wald test value of all two-way interaction variable coefficients.

The results of the analysis of the logistic regression model, which included the predictor variables of OAT-R, IOWA, ORF, race, economic disadvantage, and year of participation are listed in Table 22. The logistic regression model created from the combination of the six predictor variables explained a significant proportion of the
variance in the dependent variable, group assignment ($\chi^2 = 91.535, df = 7, p < .001$). This combination of predictor variables explained a significant percentage of the variance in the dependent variable (Cox & Snell $R^2 = .268$, Nagelkerke $R^2 = .375$).

Table 22

**Results for the Logistic Regression Model- Grade 5**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Wald test value</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OAT-R</td>
<td>-.024</td>
<td>9.453</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>IOWA</td>
<td>-.041</td>
<td>12.570</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>ORF</td>
<td>-.005</td>
<td>0.754</td>
<td>.385</td>
</tr>
<tr>
<td>Race</td>
<td>-.309</td>
<td>0.450</td>
<td>.502</td>
</tr>
<tr>
<td>Econ</td>
<td>-.640</td>
<td>2.111</td>
<td>.150</td>
</tr>
<tr>
<td>Year</td>
<td>-.421</td>
<td>1.511</td>
<td>.470</td>
</tr>
<tr>
<td>Constant</td>
<td>19.377</td>
<td>37.357</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

Step 4. The resulting logistic regression model was used to estimate a probability for each of the 293 participants in the Tier 1 and Tier 2 intervention programs. Each probability value indicated the probability that the participant would be a member of the Tier 2 intervention group. These probability values are the propensity scores for each participant. The propensity scores were used to rank the participants and to classify them into four stratified groups of approximately 73 students each. Actual group sizes were adjusted to 72, 75, 75, and 71 participants following the rule of including all identical propensity scores in the same strata (D’Agostino, 1998).
Step 5. In order to verify that the propensity score groups removed the initial bias on the covariates, a two-way ANOVA analyses was conducted on each of the covariates. For these analyses, the two main effects were the two treatment groups (Tier 1 or Tier 2 intervention) and the four propensity score groups. The probability values for each of the two-way ANOVA treatment main effects is listed in Table 21 under the heading Post-Propensity Group Formation. The results indicated that stratifying participants into four subclassifications based upon propensity scores reduced the bias on each of the covariates to nonsignificant levels. Results of the two-way ANOVA treatment main effects for each of the covariates indicated that the treatment effects were not significant for OAT-R, $F(1, 285) = 1.781, p = .183$; IOWA-R, $F(1, 285) = 1.038, p = .309$; ORF, $F(1, 285) = .317, p = .574$; Race, $F(1, 285) = .119, p = .731$, Econ, $F(1, 285) = .173, p = .678$, and year of participation, $F(1, 285) = .126, p = .723$.

Balance between Tier 1 and Tier 2 intervention groups is assumed to have been achieved when both the treatment main effects and the treatment-by-propensity group interaction effects are not statistically significant when tested for each covariate. Analysis of the treatment-by-propensity group interaction effects indicated that no significant interactions were observed for the covariates OAT-R, $F(3,285) = .117, p = .950$; IOWA-R, $F(3,285) = .113, p = .953$; ORF, $F(3,285) = 1.087, p = .355$; Race, $F(3,285) = .096, p = .962$, Econ, $F(3,285) = 1.167, p = .323$, and year of participation, $F(3,285) = 1.945, p = .123$.

Step 6. The results of independent $t$ tests of differences between Tier 1 and Tier 2 intervention groups for each of the four propensity score strata are listed in Table 23. No
statistically significant differences in reading achievement gains were observed in any of the four propensity groups. The overall treatment effect was calculated using Equation 1, and the corresponding standard error value was calculated using Equation 2. The overall estimated treatment effect was .870, which is shown in Table 23 by the difference between overall reading gain scores of 2.506 for the Tier 2 intervention group and 1.636 reading gain scores for the Tier 1 intervention group. Using Equation 2, the estimated standard error value for this estimated treatment effect was 3.612. The calculated $t$ value for the overall treatment effect, which was calculated by dividing the estimated overall treatment effect of .870 by the estimated standard error of 3.612, was $t = .241, p > .05$. The propensity score analysis indicated that the overall treatment effect on reading gain scores was not statistically significant. Students receiving the Tier 2 intervention made similar reading gains as the students in the Tier 1 reading intervention group.
Table 23

*Estimated Treatment Effects on Grade 5 Reading Gain Scores Using Propensity Score*

<table>
<thead>
<tr>
<th>Score Groups</th>
<th>Treatment</th>
<th>Group size</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>Tier 1</td>
<td>69</td>
<td>0.565</td>
<td>22.84</td>
<td>.383</td>
</tr>
<tr>
<td></td>
<td>Tier 2</td>
<td>3</td>
<td>5.667</td>
<td>9.50</td>
<td></td>
</tr>
<tr>
<td>Group 2</td>
<td>Tier 1</td>
<td>65</td>
<td>4.908</td>
<td>18.94</td>
<td>-0.789</td>
</tr>
<tr>
<td></td>
<td>Tier 2</td>
<td>10</td>
<td>-.600</td>
<td>29.61</td>
<td></td>
</tr>
<tr>
<td>Group 3</td>
<td>Tier 1</td>
<td>42</td>
<td>0.405</td>
<td>28.71</td>
<td>.610</td>
</tr>
<tr>
<td></td>
<td>Tier 2</td>
<td>33</td>
<td>3.788</td>
<td>15.52</td>
<td></td>
</tr>
<tr>
<td>Group 4</td>
<td>Tier 1</td>
<td>23</td>
<td>0.566</td>
<td>26.86</td>
<td>.096</td>
</tr>
<tr>
<td></td>
<td>Tier 2</td>
<td>48</td>
<td>1.229</td>
<td>27.59</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>Tier 1</td>
<td>199</td>
<td>1.636</td>
<td></td>
<td>.241</td>
</tr>
<tr>
<td></td>
<td>Tier 2</td>
<td>94</td>
<td>2.506</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * p<.05
Hypothesis for Research Question 2 Answered Using Propensity Score Analysis

Fifth grade students who received the Tier 2 standard protocol reading intervention will make different achievement gains on the Ohio Achievement Test in reading (p<.05) than students who received only the Tier 1 reading intervention when controlling for the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race. This hypothesis was rejected because the reading achievement gain for students receiving the Tier 2 intervention was not significantly different than the reading achievement gain of students receiving the Tier 1 intervention when controlling for the effects of prior reading achievement, oral reading fluency, economic disadvantage, and race. The treatment coefficient obtained from the propensity score analysis indicates that the estimated reading achievement gain for the Tier 2 intervention group was .870 points greater than the estimated reading achievement gain for the Tier 1 intervention group, \( t = .241, p>.05 \).

Research Question 3

3. Will there be a significant difference in the effect sizes reported from a program evaluation of the effectiveness of the Tier 2 reading intervention among third grade students when the analysis is conducted using traditional Analysis of Covariance (ANCOVA) compared to propensity score analysis?

This question was answered by first converting the Analysis of Covariance (ANCOVA) and propensity score analysis test statistics for the third grade program evaluations to a common standardized measure of effect size. Second, the effect sizes
were statistically compared to determine if there is a significant difference between the effect sizes reported by each analytic technique.

The standardized effect size calculation was computed using the formula for Hedges’ $g$ (Hedges, 1981) using the calculated $t$ values and sample sizes from the ANCOVA and propensity score analyses. The following formula was used to compute the standardized effect size value:

$$g = \frac{t \sqrt{n_1 + n_2}}{\sqrt{n_1 n_2}}$$

The standardized effect size for the ANCOVA analysis was based upon a calculated $t$ value of -3.275, and sample sizes of 188 participants in the Tier 1 intervention group and 106 participants in the Tier 2 intervention group. The resulting standardized effect size for the Tier 2 intervention based on Analysis of Covariance was -.397, which falls in the small effect size classification (Cohen, 1988).

The standardized effect size for the propensity score analysis was based upon a calculated $t$ value of -3.036, and sample sizes of 188 participants in the Tier 1 intervention group and 106 participants in the Tier 2 intervention group. The resulting standardized effect size for the Tier 2 intervention based on the propensity score analysis was-.368, which also falls in the small effect size classification (Cohen, 1988).

Statistical comparison of the two standardized effect sizes was conducted using the procedure described by Rosenthal (1991). Two standardized effect sizes can be compared using a Fisher’s $Z$ test when the estimated variance of Hedges’ $g$ is used as a measure of within-groups variance term in the denominator. The result is a single $Z$
statistic that describes the probability that the effect sizes belong to the same sampling
distribution of effect sizes. This estimate of the variance of $g$ is computed using the
following formula:

$$w = \frac{2(n_1n_2)(n_1 + n_2 - 2)}{(n_1 + n_2) \left[ t^2 + 2(n_1 + n_2 - 2) \right]}$$

A test of the significance of the difference between the two independent $g$’s can be tested using the following formula:

$$Z = \sqrt{\frac{g_A - g_B}{1 \over w_A} + \frac{1}{w_B}}$$

The estimate of the variance of Hedges’ $g$ for the ANCOVA analysis was 66.559 and the estimate of Hedges’ $g$ for the propensity score analysis was 66.729. The resulting $Z$ value testing the significance of difference between the two independent $g$’s was not significant, $Z = .167, p = .867$.

Hypothesis for Research Question 3: There will be a significant difference ($p < .05$) between the effect sizes estimating the difference between Ohio Achievement Test Reading gains for third grade students who participated in the Tier 2 intervention compared to students who received only the Tier 1 reading intervention when the differences are estimated using ANCOVA and propensity score analysis. This hypothesis was rejected because the standardized effect sizes calculated from the ANCOVA analysis and the propensity score analysis were not significantly different when tested using Fisher’s $Z$. The standardized effect size of -.397 resulting from the ANCOVA was not
significantly different from the standardized effect size of -.368 resulting from the propensity score analysis of third grade reading intervention results, \( Z = .167, p > .05 \).

**Research Question 4**

4. Will there be a significant difference in the effect sizes reported from a program evaluation of the effectiveness of the Tier 2 reading intervention among fifth grade students when the analysis is conducted using traditional Analysis of Covariance (ANCOVA) compared to propensity score analysis?

This question was answered by first converting the Analysis of Covariance (ANCOVA) and propensity score analysis test statistics for the fifth grade program evaluations to a common measure of effect size. Second, the effect sizes were statistically compared to determine if there is a significant difference between the effect sizes reported by each analytic technique. The standardized effect size calculations and differences between effect size calculations were computed using Hedges’ \( g \) and Fisher’s \( Z \) calculations as in the testing of hypotheses in Research Question 3.

The standardized effect size for the ANCOVA analysis was based upon a calculated \( t \) value of -.407, and sample sizes of 199 participants in the Tier 1 intervention group and 94 participants in the Tier 2 intervention group. The resulting standardized effect size for the Tier 2 intervention based on Analysis of Covariance was -.051, which falls in the *small* effect size classification (Cohen, 1988).

The standardized effect size for the propensity score analysis was based upon a calculated \( t \) value of .241, and sample sizes of 199 participants in the Tier 1 intervention group and 94 participants in the Tier 2 intervention group. The resulting standardized
effect size for the Tier 2 intervention based on the propensity score analysis was .030, which also falls in the *small* effect size classification (Cohen, 1988).

The estimate of the variance of Hedges’ $g$ for the ANCOVA analysis was 63.887 and the estimate of Hedges’ $g$ for the propensity score analysis was 63.837. The resulting Z value testing the significance of difference between the two independent $g$’s was not significant, $Z = -.457, p = .648$.

*Hypothesis for Research Question 4: There will be a significant difference* ($p<.05$) *between the effect sizes estimating the difference between Ohio Achievement Test Reading gains for fifth grade students who participated in the Tier 2 intervention compared to students who received only the Tier 1 reading intervention when the differences are estimated using ANCOVA and propensity score analysis.* This hypothesis was rejected because the standardized effect sizes calculated from the ANCOVA analysis and the propensity score analysis were not significantly different when tested using Fisher’s Z. The standardized effect size of -.051 resulting from the ANCOVA was not significantly different from the standardized effect size of .030 resulting from the propensity score analysis of third grade reading intervention results, $Z = -.457, p > .05$. 
CHAPTER V
DISCUSSION

The current study investigated the presence of confounding covariates such as race, economic disadvantage, prior achievement, and oral reading fluency skills in a local program evaluation of a standard-protocol Tier-2 reading intervention program. School psychologists working as systems consultants or intervention program evaluators (Ysseldyke et al., 2006) have traditionally utilized adjusted means approaches such as Analysis of Covariance (ANCOVA) or multiple regression techniques to determine the effectiveness of interventions when known confounding covariates have been present (e.g., Aarnoutse & Schellings, 2003; Balfanz, Legters, & Jordan, 2004; Gettinger and Stoiber, 2007; Modaressi & Wolanin, 2007; Vaughn et al., 2009; Wheldall, 2000; Nokes, Dole, & Hacker, 2007). Previous researchers have argued that there are significant weaknesses in the use of adjusted means approaches, most critically the undefined validity of the construct of *adjusted achievement* (Newman et al., 2005). The current study consisted of a local program evaluation of a standard-protocol Tier-2 reading intervention using the dual analyses of ANCOVA and an alternate technique, propensity score analysis (PSA) (D’Agostino, 1998; Rosenbaum, 2002; Rosenbaum & Rubin, 1983, 1984). The PSA technique made the adjustment for confounding covariates on the independent variable instead of the dependent variable, and eliminated the need for the construct of adjusted achievement (Tracz, Nelson, Newman, & Beltran, 2005). This study compared the results of these dual analyses and determined if the analyses resulted in
similar or different conclusions regarding the effectiveness of the Tier 2 intervention.

This chapter begins with a comparison of the results of ANCOVA and propensity score analysis methodologies applied to the intervention and a discussion of the practical significance of the findings. A review of the delimitations and limitations of the study is then presented. The chapter concludes with a discussion of the implications for the profession of school psychology and suggestions for future research.

Comparison of Results of ANCOVA and Propensity Score Analysis Methodologies

The purposes of this study were to determine (a) if dual analysis of Tier 2 intervention effectiveness using ANCOVA and propensity score analysis would result in similar conclusions on the treatment outcomes, and (b) if there would be a significant difference in the standardized effect sizes reported applying each methodology.

Third Grade Participants

Analysis of the data investigating potential differences in reading achievement gains using ANCOVA indicated that third grade students who participated in the Tier 2 intervention made significantly smaller gains than students participating in the Tier 1 intervention. The overall treatment effect reported for this analysis was -8.779 scaled score points on the Ohio Achievement Test in reading (OAT-R). Students who received the Tier 2 intervention for at least 27 weeks made reading achievement gains that were 8.8 points smaller than students who participated in only the Tier 1 intervention. When the same data were analyzed using propensity score analysis, the analysis similarly concluded that third grade students participating in the Tier 2 intervention made significantly smaller gains than students participating in the Tier 1 intervention. The
overall treatment effect reported for this analysis was -10.871 scaled score points on the OAT-R. Students who received the Tier 2 reading intervention for at least 27 weeks made reading achievement gains that were 10.9 points smaller than students who participated in only the Tier 1 intervention. Based on this dual analysis, the ANCOVA and propensity score analysis methodologies resulted in the same conclusion on the research question of intervention effectiveness with respect to the third grade participants.

**Fifth Grade Participants**

Analysis of the data investigating potential differences in reading achievement gains using ANCOVA indicated that fifth grade students participating in the Tier 2 intervention made the same gains as students participating in the Tier 1 intervention. The overall treatment effect reported for this analysis of -.997 scaled score points on the OAT-R was not significant. Students who received the Tier 2 intervention for at least 27 weeks made reading achievement gains that were one point smaller than students who participated in only the Tier 1 intervention, and this one point difference was not reliable beyond the threshold of chance variation. When the same data were analyzed using propensity score analysis, the analysis similarly concluded that fifth grade students participating in the Tier 2 intervention made the same reading gains as students participating in the Tier 1 intervention. The overall treatment effect reported for this analysis of .870 scaled score points on the OAT-R was not significant. Students who received the Tier 2 reading intervention for at least 27 weeks made reading achievement gains that were .870 scaled score points greater than students who participated in only the Tier 1 intervention, but this one point difference was not reliable beyond the threshold of
chance variation. Based on this dual analysis, the ANCOVA and propensity score analysis methodologies resulted in the same conclusion on the research question of intervention effectiveness with respect to the fifth grade data.

*Effect Size Comparisons*

Examination of the effect sizes reported in the ANCOVA and propensity score analysis methodologies for third grade students indicated that there was no significant difference in standardized effect sizes reported. The standardized effect size reported from the ANCOVA of -.397 was almost identical to the standardized effect size reported from the propensity score analysis of -.398. There was no significant difference in these effect sizes when statistically tested using Fisher’s Z. This means that the dual analyses not only resulted in the same conclusions with respect to the effectiveness of the Tier 2 intervention for third grade students, but also resulted in the same conclusions with respect to the *degree* of effectiveness.

Examination of the effect sizes reported in the ANCOVA and propensity score analysis methodologies for fifth grade students indicated that there was no significant difference in standardized effect sizes reported. The standardized effect size reported from the ANCOVA of -.051 was not significantly different from the standardized effect size of .030 reported from the propensity score analysis. There was no significant difference in these effect sizes when statistically tested using Fisher’s Z. This means that the dual analyses resulted in the same conclusions with respect to the effectiveness of the Tier 2 intervention and in the same conclusions with respect to the degree of effectiveness for the fifth grade students as well.
Practical Significance

Implications Regarding the Effectiveness of the Intervention

The design of this study identified the dependent variable as reading achievement gain, which was operationalized by the difference between scores on the OAT-R between two points in time. For the third grade students, the difference was computed between the OAT-R administered at the beginning of third grade and the OAT-R administered at the end of third grade. For the fifth grade students, the difference was computed between the OAT-R administered at end of fourth grade and the OAT-R administered at the end of fifth grade. The scaled scores on the OAT-R have a standard scaling for each test that establishes a scaled score of 400 as the standard of proficiency on grade level standards (Office of Assessment, Ohio Department of Education, 2009a). Because of this type of scaling, students whose scores remain constant on subsequent administrations of the test are making one year of progress on grade level standards. Students whose scores decrease on subsequent administrations of the test are making less than one year of progress, and students whose scores increase are making more than one year of progress.

Examination of the mean reading achievement gain scores of third grade students indicates that both the Tier 1 intervention group and the Tier 2 intervention group had positive gains in both the ANCOVA and propensity score analysis results. In the ANCOVA results, the adjusted reading achievement gain was 11.51 scaled score points in the Tier 2 intervention group and 20.29 points in the Tier 1 intervention group. In the propensity score analysis, the mean gain for the Tier 2 intervention group was 5.94 scaled score points while the Tier 1 group gained 16.81 points. Although the results indicated
that the Tier 1 group made significantly greater gains, the positive gain scores for each
group indicate that the students in both groups are making more than one year of progress
per school year. At the third grade level, a practical interpretation is that students in both
Tier 1 and Tier 2 interventions are increasing their reading skills at more than one grade
level per year, and students in Tier 1 are learning at a faster pace than students in Tier 2
intervention.

Examination of the mean reading achievement gain scores of fifth grade students
indicates that both the Tier 1 intervention group and the Tier 2 intervention group had
positive gains in both the ANCOVA and propensity score analysis results as well. In the
ANCOVA results, the adjusted reading achievement gain was 1.39 scaled score points in
the Tier 2 intervention group and 2.39 points in the Tier 1 intervention group. In the
propensity score analysis, the mean gain for the Tier 2 intervention group was 2.51 scaled
score points while the Tier 1 group gained 1.63 points. Although the results indicated that
both groups made similar gains, the positive gain scores for each group indicate that the
students in both groups are making at least one year of progress per school year. At the
fifth grade level, a practical interpretation is that students in both Tier 1 and Tier 2
interventions are increasing their reading skills at least one grade level per year, and
students in Tier 2 are learning at least as fast as students in Tier 1 intervention.

Implications Regarding the Appropriateness of the Methodology Used in Analysis of the

Data

The presence of a known systematic bias such as overrepresentation of African-
American students, economically disadvantaged students, or students with poor oral
reading fluency in Tier 2 intervention programs presents a critical challenge to school psychologists acting in the role of intervention program evaluators. The most common methods of controlling for this bias are the use of systematic regression techniques or Analysis of Covariance (ANCOVA). While this creates a method of controlling for the inherent selection bias, it also creates a new dilemma for the school psychologist. The result of the ANCOVA is a comparison of the adjusted means resulting from the intervention. The substitution of adjusted means for means of the original variable introduces a new construct in the results of program evaluation, a construct which may or may not have validity. If the dependent variable in a program evaluation is related to a high-stakes measurement such as the OAT-R, the results of the program evaluation are of much greater value if the results are stated in terms of the original variable.

Adjusting for known selection bias using propensity score analysis involves making the adjustment on the independent variable rather than the dependent variable. Using this technique, the results of the program evaluation are reported while both making an adjustment for the selection bias and maintaining the integrity of the construct measured in the outcome variable. In practical terms, this means that if the school psychologist wants to know how much of an effect a Tier 2 intervention will have on the OAT-R scores of students who need the intervention, the propensity score analysis technique will report the results in the same form they were originally measured. A mean reading achievement gain of 10 points can be interpreted with the expectation that most students will increase their OAT-R scores by about 10 points, the practical difference between proficiency designations of basic and proficient.
A second consequence of using ANCOVA to adjust for selection bias is the technique’s treatment of the variance contributed by the covariates responsible for the selection bias. The goal of ANCOVA is to identify the unique variance accounted for by each covariate and statistically remove the variance from the analysis. The means of treatment groups are then compared as if the effects of the covariates were not creating bias in the sample. The practical problem with this way of treating the data is that when the statistical analysis is complete, the population receiving the Tier 2 intervention is still disproportionately comprised of students who are African-American, economically disadvantaged, and have poor oral reading fluency. The school psychologist is particularly interested in knowing if the Tier 2 intervention is effective among the biased population, and is not interested in knowing if the intervention is effective regarding the biased population as if it was not (Yanovitsky et al., 2006).

A third consequence of using ANCOVA to adjust for selection bias potentially creates a more theoretical dilemma by introducing a type of error called type VI error (Newman, Deitchman, Burkholder, & Sanders, 1976; Newman et al., 2005). A Type VI error is a mismatch between the research question and the selected method of analysis. If the original challenge was to identify relatively large scale interventions that could successfully close the gap between general students and low achieving students, African-American students, or economically disadvantaged students, the method of controlling for the systematic bias removes the power of the researcher to answer the original research question. The ANCOVA technique allows the researcher to compare the gains of groups by adjusting the means to control for the selection bias. However, the original
aim of the researcher is to intentionally target students in a biasing condition and compare their unadjusted means. The use of propensity score analysis allows the school psychologist to compare the unadjusted means while still making an adjustment for the sample bias that is known to exist. The No Child Left Behind Act (NCLB, 2001) specifically requires local school districts to demonstrate AYP among the general population and specific groups of students including African-American students, students with disabilities, and English language learners (State Board of Education Closing Achievement Gaps Task Force, 2003). It is of critical importance to identify those interventions that are successful in helping schools close these achievement gaps through the most effective research methods. The use of the propensity score analysis technique may be a useful tool in determining if achievement gaps are being closed specifically among these subgroups.

Delimitations of the Study

There were several important factors that served as delimitations to this study. These factors are the use of a quasi-experimental design and the exclusion of participants.

Quasi-Experimental Design

Regarding this investigation of the effectiveness of the Tier 2 reading intervention, the most significant delimitation was the use of a quasi-experimental research design. A true experimental design would have required both random of selection of students to be included in the study as well as random assignment of subjects to Tier 1 and Tier 2 intervention groups. For the purposes of this study, students were included because they belonged to an intact group (i.e., all third grade and fifth grade
students in a suburban district’s only intermediate elementary school). Instead of being randomly assigned to treatment groups, the students were assigned to treatment based upon eligibility criteria (i.e., DIBELS measures of oral reading fluency). The consequence of this non-random assignment of participants is that differences in the outcome of the groups on the dependent variable (i.e., reading gains) cannot be unequivocally attributed to changes in the independent variable (i.e., treatment).

Despite the critical advantages of employing a true experimental design, there were factors present in this study that would caution against the use of experimental design. A particular consideration was the ethical issue of the standard of care (Shadish et al., 2002). The development of reading skills in elementary school can place students on a trajectory toward future academic success or failure. It would not have been ethical to deprive the students with the lowest reading skills the opportunity to receive more intensive intervention aimed at improving their reading trajectory. In addition, because this study was conducted in an applied setting, the allocation of resources necessary for the intervention required financial accountability. Because the cost of the Tier 2 intervention limited its availability to 20% of the population, and the outcome variable related to a high-stakes test that determined state-level designations, some form of systematic assignment of participants was necessary. As such, eligibility was a consideration based upon the non-random criterion of oral reading fluency skills.

Exclusions of Potential Participants

Another delimitation in this study was the extent to which a very large number of potential participants could not be included in the final data set for analysis. Of the
original 573 third grade students in attendance during the 2005-2008 school years, 294 were ultimately included in the study. Therefore, 48.7% of the potential participants were excluded based upon factors such as transience, attendance patterns that caused missing test scores, and incomplete documentation. A defining characteristic of the participants included in the analysis was consistency in attendance, as all of these students were present for all assessments and were residents of the school district for at least one school year. This was necessary for all analyses.

Of the original 579 fifth grade students in attendance during the 2005-2008 school years, 293 were ultimately included in the study. The exclusion rate for the fifth grade group was 49.4%, and was also primarily affected by the transience of students who were not present for one full school year. This criterion was not set arbitrarily, but was the result of the use of gain scores on the Ohio Achievement test in reading. Students had to be present for two consecutive annual administrations of the test in order for gain scores to be computed.

An unexpected delimitation related to exclusion of participants was the overall percentage of participants who received the Tier 2 intervention for the full 27 weeks. Although approximately 20% of the overall grade level population was provided with the Tier 2 intervention at any given time, the percentage of the final participant group (i.e., after exclusion of participants) comprised 36.1% of the third grade study population and 32.1% of the fifth grade study population. This means that there were more students in the Tier 1 intervention group excluded from the final participant group than from the Tier 2 intervention group, suggesting that transience of the population could be affecting the
results in ways that were not anticipated. Because participation in the Tier 2 intervention was based upon meeting eligibility criteria and participating for at least 27 weeks, it may have been that the Tier 2 group had an overall better attendance pattern than the Tier 1 group. This would be an important factor for study because students with better attendance would be expected to make greater reading gains simply because they were in school more consistently. Because this study was based on an existing data set, it was not possible to investigate the attendance patterns of the participants in retrospect.

Limitations of the Study

This study had several limitations to consider upon drawing conclusions related to implication for research and practice. They are identified and discussed in this section.

Limitations with Respect to Treatment Fidelity

An important limitation to the study is related to treatment fidelity, the extent to which an intervention is implemented as intended (Gresham, Gansle, Noell, Cohen, & Rosenblum, 1993; Telzrow & Beebe, 2002). Because this study involved a program evaluation based on an existing data set, the procedures in place to ensure that the interventions were implemented as intended were somewhat limited. Nevertheless, there were factors that provided some control of treatment integrity. First, all of the adults providing the intervention were licensed teachers with training in primary reading instruction. Second, all of the teachers were provided with several days of uniform training in the summer prior to the school year to implement the program. Third, all of the teachers providing the intervention used identical instructional manuals throughout the school year. Fourth, all of the teachers providing the intervention continued to
participate in weekly collegial planning sessions in which they shared planning
information about weekly lessons. Last, the grade-level building configuration allowed
all of the teachers providing the interventions to work in the same building and
participate in the weekly collegial planning.

Despite these positive contributions to treatment fidelity, there were significant
limitations. First, the interventions were implemented in small groups, by several
teachers, over two grade levels, across a three year period. During that time, there was
significant turnover in part-time staff from year to year. Although none of the staff
turned over during any of the three school years, there was substantial turnover between
school years. As a percentage of intervention staff turnover from one year to the next,
there was a 50% turnover in staff in 2005-2006, a 75% turnover in 2006-2007, and a 25%
turnover in 2007-2008. As a result, grade level groups were often taught by different
teachers from year to year and these teachers varied in their level of experience teaching
the intervention. This inconsistency in teachers providing the intervention could have
negatively affected the fidelity of the intervention. This may have been the case
particularly among fifth grade students in the study, as the covariate for year of
participation was included in the multiple linear regression model, indicating that year of
participation explained a unique and significant proportion of the variance in reading
achievement gains among fifth grade students. The results from the third grade program
evaluation did not reveal any significant confounder due to year of participation.

In order to know the degree to which an intervention is implemented as designed,
it is necessary to have tools for measuring treatment fidelity. Methods that have been
reported in the literature include participant reports, evaluation of permanent products, and collecting reports from outside sources (Telzrow & Beebe, 2002). In this study the participant reports were provided weekly by teachers, but were informal reports given orally in collegial planning meetings only. These reports, however, should not be considered highly valid measures of treatment integrity, as self report data have been shown to be highly unreliable in the literature (Wickstrom, Jones, LaFleur, & Witt, 1998). The most valuable data for integrity checks are reports collected from outside sources. The preferred method for these integrity checks are incorporating direct observations of intervention events by trained observers across multiple trials and settings (Moncher & Prinz, 1991). This type of procedure was not incorporated in the study and may have provided critical information about the validity of the study.

**Limitations with Respect to Instrumentation**

**DIBELS**

The Dynamic Indicators of Basic Early Literacy (DIBELS; Kaminski & Good, 1996) is a commonly utilized, curriculum-based benchmark assessment in reading, one of many tools available to school psychologists for progress monitoring (Burns & Coolong-Chaffin, 2006; Brown-Chidsey & Steege, 2005). As a progress monitoring instrument, DIBELS has been demonstrated to produce highly reliable and valid results. Poor response to the universal intervention is typically operationalized as failure to meet the benchmark criteria at grade level identified by the curriculum-based measure (Jenkins, Hudson, & Johnson, 2007; Burns & Coolong-Chaffin, 2006). The benchmark criteria are
based on both national normative criteria and local norms for the average performance of students in the same grade at the child’s school (Brown-Chidsey & Steege, 2005).

In the current study, students with the lowest 20% of DIBELS oral reading fluency scores for the grade level were assigned to Tier 2 intervention, a practice recommended by researchers (Burns et al., 2007; Brown-Chidsey & Steege, 2005). However, this procedure resulted in varying absolute criteria for assignment to the Tier 2 treatment condition from one year to the next. The criterion cutoff score for eligibility that identified the lowest 20% of readers differed from year to year as the local norms changed in each grade-level population. This limitation affected the assignment of participants to the Tier 2 intervention and could therefore have affected their responsiveness to intervention. Prior research has shown that measures of oral reading fluency have moderate to strong correlations with statewide proficiency test results (Baker et al., 2008; Barger, 2003; Wilson, 2005) and with the Ohio Reading Proficiency Tests, in particular (Vander Meer, Lentz, & Stollar, 2005). Since the outcome variable for this study (i.e., reading achievement gain) was based on proficiency test score changes, there may have been a differential response to the intervention based on the eligibility cutoff scores used in each of the year during which participant performance was measured and analyzed.

*Ohio Achievement Test – Reading (OAT-R)*

Published technical reports obtained from the Ohio Department of Education regarding the validity and reliability of OAT-R scaled scores indicated satisfactory reliability coefficients of .86 to .88 (Office of Assessment, Ohio Department of
Education, 2009a). However, it could not be determined whether this was a measure of internal consistency reliability, test-retest reliability or alternate forms reliability. Because the OAT-R was used as a measure of baseline reading achievement, the preferred measure of reliability was the test-retest type. A test with high internal consistency does not necessarily yield consistent scores from one administration to the next. This uncertainty of information about the baseline reading achievement score data should indicate caution when interpreting the measure of the dependent variable as well as the unique variance explained by the baseline reading achievement level.

Limitations Concerning Reliability and Standard Error of Measurement

The dependent variable in this study was the measure of reading achievement gain as determined by the difference between consecutive administrations of the OAT-R. Because this study involved administration of the OAT-R in two consecutive years or a fall and spring administration as the basis of the dependent variable, it would have been helpful to have technical information regarding test-retest reliability in particular. Differences in scores from one administration to the next could have been interpreted with more clarity given a specific measure of test-retest reliability.

Standard errors of measurement of 9.67 scaled score points for the third grade test and 10.73 scaled score points for the fifth grade test were reported in the technical data for the OAT-R (Office of Assessment, Ohio Department of Education, 2009a). Again, it could not be determined if these standard errors of measurement are based on measures of internal consistency reliability or test-retest reliability. These standard error values are important to the current study, because the magnitude of difference in gain scores for
the Tier 1 and Tier 2 intervention groups was generally within the standard error of measurement of the test instrument. For example, in the third grade study, the ANCOVA analysis indicated an adjusted gain of -8.78 scaled score points for the Tier 2 group and the PSA analysis indicated a -10.87 scaled score point difference. Only the result of the PSA analysis demonstrated a difference in gain scores greater than the standard error of the test instrument. Therefore, the results should be interpreted in light of this measurement limitation.

**Limitations Concerning Scaled Score Confidence Intervals**

The use of gain scores as the dependent variable introduced a potential compounding effect on the standard error of measurement confidence intervals. Since each administration of the OAT-R resulted in an obtained score surrounded by a standard error band, the difference between the scaled scores from two administrations of the OAT-R could have a standard error band that is a multiple of the reported standard error band when no difference in scores exists. For example, the standard error of measurement is reported as 10.73 scaled score points on the fifth grade test. A student scoring 410 on the baseline administration would have a standard error band of approximately 400 to 420 at the 68% confidence interval. If the student’s score on the post-test administration was within the 68% confidence interval at a scaled score of 420, that student’s 68% confidence interval for the second administration would be from 410 to 430. Therefore, this student who scored within the standard error band on sequential administrations of the OAT-R could have a combined error band between 400 and 430 scaled scores on the new variable *reading achievement gain*. This compounding effect of
the standard error of measurement should be considered a potential limitation in interpreting any reported reading gain scores as well as any reported group differences in reading gain scores.

Limitations Concerning Sensitivity of Instrumentation

In addition to these concerns, the OAT-R may not be an appropriate measure in determining subtle changes in academic achievement over a long period of time if only two samples are conducted. Many different factors such as motivation or fatigue could possibly impact a student's performance on the OAT-R on any one particular administration date. The design of the study limited the measure of academic achievement over time to only two OAT-R data points. Thus, potential confounds on either of those two data points (e.g., lack of proper breakfast, poor night’s sleep, illness on the date of the examination) might seriously detract from the validity of a particular student’s estimated reading achievement gain.

Limitations Concerning the Timing of Test Administrations

An additional concern in using the OAT-R as a measure of reading achievement to determine reading achievement gains was related to the timing of the administrations of these tests. At the third grade level, the baseline administration occurred in October, after the students had received up to 7 weeks of either the Tier 1 or Tier 2 interventions. Also at that grade level, the spring administration of the OAT-R was given in different months each year due to changes in the state proficiency testing legislation. For the 2005-2006 cohort, the outcome measure was administered in March; for the 2006-2007 and 2007-2008 cohorts the outcome measure was administered in May. Therefore, the
number of months between baseline measure and outcome measure was 5 months for the 2005-2006 cohort and 7 months for the 2006-2007 and 2007-2008 cohorts.

The fifth grade reading achievement gain scores were computed from the difference between the scores on the fourth grade OAT-R and the fifth grade OAT-R. Changes to the state legislation governing proficiency testing also impacted the timing of these assessments, resulting in different test intervals for each cohort. For the 2005-2006 cohort, the baseline measure was given in March, 2005 and the outcome measure given in March, 2006. For the 2006-2007 cohort, the baseline measure was given in March, 2006 and the outcome measure given in May, 2007. For the 2007-2008 cohort, the baseline measure was given in May, 2007 and the outcome measure was given in May, 2008. As a result of these differences in timing, the test interval for the 2005-2006 cohort was 12 months, for the 2006-2007 was 14 months, and for the 2007-2008 cohort was 12 months. These differences in test intervals and timing of test administrations during the academic year may have significantly impacted the comparability of gain scores between cohorts, and may be responsible for the identification of a significant covariate for year of participation in the fifth grade study.

*Limitations Concerning Content Validity of the Dependent Variable*

The Ohio Achievement Tests are criterion-related tests based on grade level content standards established by the Ohio Department of Education (Office of Assessment, Ohio Department of Education, 2009a). Because the content standards are designed to increase in complexity with each successive grade level, the achievement tests designed to measure proficiency in those content standards reflect the changes in
those standards. The implication for the current study is that within the third grade study, both the baseline measure and the outcome measure were based strictly on third grade content standards. However, the baseline measure used in the fifth grade study was based on fourth grade standards while the outcome measure was based on fifth grade standards. Therefore, the two assessments were not measuring precisely the same constructs at the fifth grade level, while the two assessments given in the third grade study were designed to measure precisely the same constructs. This limitation impacts the generalization of the findings, especially of the fifth grade study, and may be one explanation that students in the fifth grade study showed lower gain scores overall in both the Tier 1 and Tier 2 intervention groups.

**Potential Multicollinearity of Variables**

Pedhazur (1997) described a problem of interpreting results of a multiple regression analysis when the intercorrelations between independent variables are high, called multicollinearity. Three potential problems that can result from multicollinearity of independent variables are (a) the condition significantly limits the size of $R$, the maximized multiple correlation, (b) the importance of individual predictors is difficult to determine because of the variance shared with other predictors, and (c) the resulting prediction equation becomes less stable (Stevens, 2007). One method of identifying multicollinearity is simple examination of the correlation matrix of predictor variables. Using this approach, the researcher identifies predictor variables with high intercorrelations that may be explaining the same variance in the regression equation. In this case, the researcher may choose to eliminate one or more variables. Another method
of identifying multicollinearity is using the statistical measure of variance inflation factor (VIF). The VIF indicates whether there is a strong linear relationship between a predictor and the predictors not yet entered in the regression model (Stevens, 2007). Generally, VIF values greater than 10 should alert the researcher that multicollinearity may be a detriment.

In the current study of third grade students, several intercorrelations were greater than $r = .55$ (ORF and IOWA, ORF and OAT-R, IOWA and OAT-R, Econ and Race). Of these paired correlations, only the pair of variables IOWA and OAT-R both appeared in the resulting regression model equation. Additionally, the VIF index for both of these variables was approximately 3.1, which is well under the rule of thumb recommended by Stevens. Ultimately the decision was made to include both variables in the model because they are conceptually measuring different, though related constructs. The OAT-R is a criterion-referenced measure of grade level reading proficiency standards, and the IOWA is a norm-referenced measure of cross-grade level reading skills. Nevertheless, the size of the intercorrelations between these variables may warrant cautious interpretation of the model.

In the current study of fifth grade students, several intercorrelations greater than $r = .55$ (ORF and IOWA, ORF and OAT-R, IOWA and OAT-R) were also observed. Of these paired equations, all variables appeared in the resulting regression model equation. The VIF indexes for these variables were approximately 2.0 in all cases, which is well below the rule of thumb recommended by Stevens (2007). Again, the ultimate decision to include these variables resulted from conceptual considerations.
inclusion or exclusion of one or more of these variables did not seem to increase the ability of the resulting regression model to detect any additional effect of the treatment. Nevertheless, the size of the intercorrelations between these variables should also be a cause for cautious interpretation of the model.

**Limitation Related to Sample Size**

The use of ANCOVA to estimate differences in adjusted scores when known covariates are creating a bias in the sample requires a recommended sample size that is present in most school settings. In an elementary school with 125 students per grade level and 20% of the population assigned to Tier 2 intervention, approximately 100 students would receive Tier 1 intervention and 25 students would receive Tier 2 intervention. This sample size would be of reasonable size to analyze the data using ANCOVA and have reasonable power to detect differences that may exist (Huitema, 1980; Pedhazur, 1997; Stevens, 2007). The current study involved a school of similar student body size, although high rates of transience required a high rate of subject exclusion.

The use of propensity score analysis to estimate differences when known covariates are creating a bias in the sample requires a significantly larger sample size than is present in most school settings. The sample size needed must take into consideration that four or five strata are created as subclassifications of the original participant sample, requiring sample sizes of approximately 250 participants (D’Agostino, 1998; Fraas et al., 2006; Yanovitsky, Zanutto, & Hornik, 2005) The size of sample required to ensure sufficient statistical power needs to be 250 or more because each subclassification should
contain a minimum of 50 to 60 subjects, and each subclassification compares Tier 1 and Tier 2 intervention groups. This poses a significant limitation for most school psychologists conducting local program evaluations. The current study attempted to compensate for a smaller student body by pooling the data from students who participated in the interventions over a three year period. However, this attempt to compensate for sample size limitations introduced new potential covariates (i.e., year of participation) and challenges of internal control (i.e., treatment fidelity) discussed in previous sections.

An inspection of the composition of subclassifications created by propensity score analysis in the current study indicates that although the number of subjects assigned to each stratum was essentially equal, the balance of Tier 1 and Tier 2 intervention recipients was unbalanced within each stratum. This condition was anticipated as described in the propensity score analysis literature (D’Agostino, 1998; Rosenbaum & Rubin, 1983; Yanovitsky et al., 2005). For example, Table 16 shows that the lowest two strata in the third grade analysis each contained only 5 subjects who received the Tier 2 intervention. Similarly, Table 23 shows that the first stratum contained only 3 subjects in the Tier 2 intervention and the second stratum contained 10 subjects in Tier 2. In this instance, the researcher must make a determination whether the subjects in these strata are contributing to the analysis. Examples in the literature suggest that if a stratum contains no subjects in a treatment condition, the subjects in that stratum are not contributing to the analysis and the researcher may decide to exclude such cases from the study (Hahs-Vaughn & Onwuegbuzie, 2006). In the current study, the numbers of
subjects in these strata were greater than zero and it was decided that data from these
subjects were contributing to the analysis. Nonequivalent treatment group sizes within
each stratum were addressed statistically using the independent t test statistic for unequal
sample sizes to test the significance of any differences in each stratum. This statistic
makes adjustment for the difference in treatment group sizes in the error term of the
computation.

Implications for the Profession of School Psychology

In the next decade, school psychologists are envisioned playing an increasing role
in addressing the academic achievement of students with and without disabilities in the
general education environment (Burns, 2007; Burns et al., 2006; Ysseldyke et al., 2006).
The Blueprint for Training and Practice III (Ysseldyke et al., 2006) identifies those
functional and behavioral competencies within the role and function of school
psychologists as they are engaged in practice at multiple levels in the intervention
process. Blueprint III suggests that school psychologists should be working towards (a)
improving competencies for all students, and (b) building and maintaining the capacities
of systems to meet the needs of all students. Specific functional competencies promoted
by Blueprint III that are relevant to the task of closing the achievement gap include
enhancing the development of cognitive and academic skills in all students, systems-
based service delivery, and data-based decision making and accountability.

Accomplishing these two outcome goals suggests a new vision for the field of
school psychology. This vision of an emerging role for school psychologists places a
premium value on not only designing but also conducting evaluation of both individual
intervention programs and standardized programs designed for implementation among
groups of students with and without disabilities. The role envisioned for today’s school
psychologist in *Blueprint III* demands that practitioners in the field move far beyond
knowledge about research methodology skills to applying these skills in conducting
program evaluations of the systems of intervention put into place for both individual
students and groups of students. They have also received extensive training in statistics,
research methods, and program evaluation, and it is this training that enables them to be
data-based decision makers on behalf of school systems, programs, and classroom
environments (National Association of School Psychologists, 2008).

The application of propensity score analysis methodology to evaluations of Tier 2
intervention programs can allow school psychologists to appropriately address the
underlying questions facing school systems. This methodology allows school
psychologists to determine if the targeted interventions are effective with the type of
students attending their schools, including many of the biases inherent in the local
demographics and the biases inherent in applied eligibility criteria. Furthermore,
propensity score analysis can offer the added practical benefit of reporting treatment
effects in their original construct (i.e., not relying on *adjusted* outcome variables). This
can allow school psychologists to make practical predictions about the impact Tier 2
interventions may make on students’ trajectory toward closing achievement gaps and
performing at proficient levels on state assessments.
Implications for Future Research

There are a number of lines of future research that may further illuminate the findings in this study. Future investigations of the effectiveness of Tier 2 reading interventions may utilize more sensitive measures of the dependent variable. Because the standards-based, state designed assessments are the basis of AYP determinations, evaluation of intervention programs use scores on these high stakes state assessments as the criterion variable in classifying interventions as effective. Other measures of student progress such as curriculum-based measurement procedures are available to researchers as measures of intervention effectiveness (Deno, 1985; Fuchs & Fuchs, 2002; Shinn, 1989, 2002). In particular, measures of oral reading fluency have been demonstrated as sensitive to small changes in skill acquisition and efficient in administration to be used in progress monitoring (Baker et al., 2008; Chidsey-Brown & Steege, 2005; Shinn, 1998). Using measures that are more sensitive to small changes in reading skills may improve the power of the study to detect differences in treatment effectiveness.

Another line of investigation that may be useful is an application of propensity score analysis to program evaluation of Tier 2 interventions in a very large school district. As one of the limitations of this method is the substantial sample size needed, the advantage of using a larger district would be to eliminate the necessity of pooling data across more than one year of participation. Pooling of data was a limitation to the current study because it introduced a new covariate (i.e., year of participation) and reduced the ability of the researcher to internally control for variables such as consistency of teaching.
staff. Application of this technique to a study of Tier 2 interventions in a large school
district may improve the power to detect outcome differences.

The current study could have been significantly improved by utilizing direct
measures of treatment fidelity. Future investigations of the effectiveness of Tier 2
interventions could incorporate systematic observations of instructional delivery and
measures of inter-observer reliability such as those described by Moncher & Prinz (1991)
to ensure that the intervention is delivered consistently as designed. This may
significantly affect the power of the study to detect improvements in some interventions
over others.

The introduction of a previously overlooked data analysis technique such as
propensity score analysis to another field of study may allow researchers to conceptualize
previously investigated research questions through the lens of the new analytic technique.
This may allow researchers to create a better match between the research question and the
analytic technique, thus avoiding Type VI errors. School psychologists routinely interact
with research questions that apply to biased participant populations. For example,
students with disabilities are disproportionately represented in the juvenile justice system.
African-American students are disproportionately represented in programs for students
with severe emotional disturbance. Boys are disproportionately represented in programs
for students with significant reading delays and in programs for students with emotional
disturbance (State Board of Education Closing Achievement Gaps Task Force, 2003).
The propensity score analysis methodology may offer a different analytic technique that
avoids many of the problems inherent to traditional methods of controlling for systematic
bias. In particular, the outcome variables retain their original construct and have advantages for interpreting the practical significance of the findings.
REFERENCES
REFERENCES


presented at the meeting of the Midwestern Political Science Association, Chicago, IL.


Barker, T., & Torgesen, J. K. (1995). An evaluation of computer-assisted instruction in


learning disabilities: Introduction to the special series. *Assessment for Effective Intervention, 32*(1), 3-5.


Exceptional Children, 49, 36-45.


academic achievement across twenty-nine high and low income countries.


Paper presented at the meeting of the Midwestern Political Science Association, Chicago, IL.


Lieberman, E., Lang, J. M., Cohen, A., D’Agostino, R., Datta, S., & Frigoletto, F. D.
*Obstetrics and Gynecology, 88*, 993-1000.

Indianapolis, IN: Dog Ear Publishing.


*Psychological Methods, 9*(4), 403-425.

*School Psychology Review, 33*(2), 193-204.


Myers, W. O., Schaff, H. V., Gersh, B. J., Fisher, L. D., Kosinski, A. S., Mock, M. B.,


critical race theory-disability studies: A robust approach to reducing racial
discrimination in evaluation decisions. In S. Jimerson, M. Burns, & A.
VanDerHeyden (Eds.), Handbook of response to intervention: The science and
practice of assessment and intervention (pp. 65-79). New York: Springer.


Newman, I., Deitchman, R., Burkholder, & J., Sanders, R. (1976). Type VI error:
Inconsistency between the statistical procedure and the research question.
Multiple Linear Regression Viewpoints, 6(4), 1-19.

analytical technique with the research question: A discussion of current research
issues. Paper presented at the annual meeting of the Eastern Educational
Research Association, Sarasota, FL.


heuristics while reading historical texts. Journal of Educational Psychology,
99(3), 492-504.

Novak, S., Reardon, S., Raudenbush, S., & Buka, S. (2006). Retail tobacco outlet
density and youth cigarette smoking: A propensity-modeling approach. American


Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and


Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 3-20). Bethesda, MD: National Association of School Psychologists.


Shapiro, E. S. (2004). *Academic skills problems: Direct assessment and intervention* (3rd


of School Psychologists, Bethesda MD.


The academic and political implications of Type VI errors in studies of achievement and socioeconomic status. *Multiple Linear Regression Viewpoints*, 31(1), 19-24.


University of Oregon Center on Teaching and Learning. *DIBELS Data System*. Retrieved January 3, 2009, from [https://dibels.uoregon.edu/](https://dibels.uoregon.edu/)


Meeting the highly qualified teachers challenge: The Secretary’s second annual report on teacher quality. Washington, DC: Author.


a structured arts program on the psychosocial functioning of youth from low-income communities: Findings from a Canadian longitudinal study. *Journal of Early Adolescence, 26*(2), 186-205.


