2016

The Effects Of Homogeneous Differentiation On Reading Achievement: Within-Class Grouping Versus Between-Class Grouping

Sharon Suzanne Liddell
University of Mississippi

THE EFFECTS OF HOMOGENEOUS DIFFERENTIATION ON READING ACHIEVEMENT:

WITHIN-CLASS GROUPING VERSUS BETWEEN-CLASS GROUPING

A Dissertation
presented in partial fulfillment of requirements
for the degree of Doctor of Philosophy
The University of Mississippi

by

S. SUZANNE LIDDELL

May 2016

ABSTRACT

The purpose of this quasi-experimental, ex-post facto study was to determine through quantitative analysis of longitudinal data, the effects of homogeneous differentiation on reading achievement. Specifically, the study sought to determine the achievement differences in reading of students who were taught using a within-class grouping method or a between-class grouping method. The study population consisted on students from a Mississippi school district who began first grade from the 2006-2007 school year through the 2009-2010 school year. Students who had been assessed over a three-year on the DIBELS, STAR, and the MCT2 assessment were eligible for inclusion in the study. A refined population sample of 240 subjects was identified. The data was analyzed using a multivariate analysis of variance and a repeated measures analysis. The results of the MANOVA indicated a significant statistical difference in achievement based on instructional grouping format. The results of the repeated measures analysis revealed no significant statistical differences in grouping format over time.

DEDICATION

This dissertation is dedicated with my deepest love and appreciation
to my late father
Vernon Brown
who taught me to persevere
and
to my mother
Jean White Brown
who taught me everything else.

ACKNOWLEDGMENTS

I have spent a lifetime as a student. I have always loved school and always loved to learn. It has only been possible for me to complete my education with the help of many wonderful people who all happen to be teachers in one way or another. Thank you for advising me, educating me, and loving me along the way. Thank God for placing you in my life!

I humbly thank my doctoral program advisor and dissertation chair, Dr. Dennis Bunch. He gave so generously of his time, held me accountable, and encouraged me to stay the course. There is absolutely no way I would have finished my dissertation without him and I am forever grateful for the profound contribution he has made to my education.

I would also like to express my gratitude to Dr. Doug Davis, Dr. Kerry Holmes, and Dr. Ryan Niemeyer for serving on my dissertation committee. I sincerely appreciate their commitment to my doctoral work and thank them for being a part of my journey.

Many thanks to my friends and colleagues who listened without judgment and helped me survive, especially Dr. Martha McLarty and Jennifer Sullivan.

I would like to thank my siblings Angela, Rebecca, and Joel who were my first teachers. They read to me, taught me to write my name, and helped me practice my debating skills almost daily. I thank them for allowing me to be their baby sister and loving me anyway.

Finally, with all my love, I acknowledge my husband Todd Liddell. He encouraged me to pursue my goals and never complained about coming in second to my work or my education. He supported me when I was confident and determined and he was my shoulder to cry on when I was ready to give up. I thank him for believing in me and reminding me to believe in myself!

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER ONE

Introduction

The enactment of Public Law 107-110, commonly known as No Child Left Behind (NCLB) a revision of the 1962 legislation known as the Elementary and Secondary Schools Act (ESEA), created an extraordinary level of accountability in K-12 public education. As part of NCLB, schools are required to show proficiency and yearly academic growth in the areas of reading and mathematics (No Child Left Behind Act of 2001). When NCLB was enacted, the original law was clear, by 2014, schools failing to meet their achievement goals would face sanctions such as reconstitution of their schools with new teachers, administrative take-over by the state, and outright school closure (Daly, 2009). As the 2014 achievement deadline approached, it became evident many schools across the country had not met the goals of NCLB and faced serious consequences.

In an effort to provide schools the opportunity to continue working toward the goal of 100% proficiency among students, The United States Department of Education (USDE) allowed states to submit a NCLB/ESEA waiver beginning in 2011, thus removing the threat of immediate sanctions and allowing schools more time to make progress. According to the U. S. Department of Education (n.d.), 43 states have submitted and received approval for their waivers as of 2015. Nonetheless, there are still serious implications for schools failing to meet achievement goals in reading and mathematics. These implications include a loss of federal funding, termination of principals, and take over of districts by the state when schools fail to show adequate yearly progress. With this threat looming, school leaders seek to implement research-based best

practices specifically in regard to reading achievement (Chorzempa & Graham, 2006).

One focus area of best practices research in reading is differentiated instruction. Differentiated instruction has been a popular topic in education since the early 1990s. Noted educator, Carol Ann Tomlinson, is often credited with coining the term and defining the concept of differentiation. Today differentiation is generally regarded as a way in which the teacher varies instruction for individual students or small groups of students based on interest, learning profile, or rate of learning in order to create the best learning environment possible for student achievement (Tomlinson, 2000). Because differentiated instruction is focused on meeting the needs of all students and producing high levels of achievement, the concept has been researched and often touted as a best practice model of instruction. For example, in 2009, the Institute of Education Sciences declared differentiated reading instruction to be a best practice for students at all instructional tier levels.

Because positive achievement outcomes have been noted with differentiated instruction, teachers and instructional leaders have chosen to implement this method in their classrooms and schools. Likewise, many state educational agencies across the nation include differentiation as part of their teacher appraisal instrument. For example, as part of the Texas Teacher Evaluation and Support System (2014), the Mississippi Statewide Teacher Appraisal Rubric (2014), and the Massachusetts Model System for Teacher Evaluation (2012), teachers are required to show proficiency in the area of differentiated instruction, thereby suggesting differentiation is expected to impact instruction and improve student achievement.

Differentiated instruction can be accomplished by grouping students according to rate of learning. There are two basic grouping models (Huebner, 2010). These are the heterogeneous model and the homogeneous model. First, in the heterogeneous model, students of various skill

or ability levels are placed in groups and learn together. Such is the case with cooperative learning groups and peer assisted learning strategies. Secondly, in the homogenous model, students of the same skill or ability level are placed in groups and learn together. Such is the case with leveled-reading groups. Of these two models, homogeneous grouping seems to be more prevalent. In a recent survey, approximately 60% of teachers responded they group students homogeneously for classroom instruction (Sparks, 2013).

Research into the effects of homogeneous grouping spans a 30-year period (Abadzi, 1984; Chueng, & Rudowicz, 2003; Condron, 2008; Gentry & Owen, 1999, Hong & Hong, 2009; Kerckhoff, 1986; Kulik & Kulik, 1992; McCoach, O'Connell & Levitt, 2006; Meijnen & Guldemond, 2002; Rowan & Miracle, 1983). Typically these studies have focused on the achievement outcome differences between students. One methodology has been to examine the achievement of students who have been placed in ability groups based on Intelligence Quotient (IQ) scores. Examples of these include Chueng and Rudowicz (2003) and Preckel, Gotz, and Frenzel (2010). Another methodology has been to examine the achievement of students who have been placed in skill-level groups based on achievement test scores. Examples of these studies include Rowan and Miracle (1983), Hong and Hong (2009), and Condron (2008).

There are two fundamental grouping formats used in homogeneous differentiated instruction (Slavin, 1987; Kulik and Kulik, 1992). These are within-class grouping and between-class grouping. Within-class grouping is a format in which students of a similar level are instructed in small homogeneous groups within a heterogeneous classroom. Between-class grouping is a format in which students of a similar level are taken from a heterogeneous classroom, placed in a homogeneous classroom, and instructed with same level peers for a portion of the day.

The results of the most recent studies of homogenous differentiation have shown mixed results. Some have found homogenous grouping to be conducive to student achievement across various skill levels (Hong & Hong, 2009; Rogers, 2007). However, others have found homogeneous grouping to be conducive to student achievement for high-level learners, but detrimental to the achievement of low-level learners (Condron, 2008; McCoach, O'Connell & Levitt, 2006). In spite of the number of studies conducted over the decades in regard to homogeneous grouping and the number of times the topic has been analyzed, the studies have often been limited by small sample sizes and one year of achievement data based on a single assessment. Consequently, many research results cannot be generalized to a greater population and the decision to differentiate or group students homogenously for instruction remains a personal and professional choice of the teacher or instructional leader.

Overall, studies regarding homogeneous grouping and student achievement tend to focus on ability or IQ grouping rather than skill-level grouping, do not focus primarily on reading achievement, and usually measure achievement based on the use of one assessment. Furthermore, studies concerning homogeneous grouping often take into account only one group of students in one grade level and simply measure the effect of grouping based on one year of student achievement data. Generalizations have been made, but overall the research fails to clearly show if a best instructional practice for homogeneous grouping exists. Therefore, a significant gap in the research is present and should be addressed.

Statement of the Problem

The purpose of this quasi-experimental, ex-post facto study is to determine through quantitative analysis of longitudinal data, the effects of homogeneous differentiation on reading achievement. Specifically, the study seeks to determine the effects of within-class skill level

grouping versus the effects of between-class skill level grouping on student achievement in reading. Student scores on the Dynamic Indicators of Early Literacy Skills (DIBELS), the Standardized Test for the Assessment of Reading (STAR), and the Mississippi Curriculum Test 2nd Edition (MCT2) Language Arts subtest are analyzed to determine if there are statistical differences between the scores based on grouping format.

Significance of the Study

In the age of high-stakes testing and legislation such as No Child Left Behind, achievement outcomes are of extreme importance and impact a school's ability to remain operational. School leaders are charged with the task of overseeing the instructional process, analyzing student data for growth and achievement outcomes, and making instructional program choices most conducive to student learning. Based on the idea that principal leadership practices including instructional planning and development have a great influence on student achievement, (Denton, Foorman & Mathes, 2003; O'Connell & White, 2005; Reitzug, West & Angel, 2008), programming decisions are a primary responsibility of school leaders. If particular instructional grouping formats have the potential to positively affect achievement outcomes, school leaders should be made aware of them so they are better able to implement practices having the greatest impact on overall achievement. Current research indicates the need for homogeneous differentiated instruction but fails to clearly suggest which format of homogenous grouping has the most beneficial outcome on student achievement. The study proposed herein is designed to address two forms of homogeneous differentiated instruction: within-class skill level grouping and between-class skill level grouping. Within-class grouping is a structure in which students of the same skill level are instructed in small homogeneous groups within a heterogeneous classroom. Between-class grouping is structure in which students of the same skill level are

taken from a heterogeneous classroom, regrouped to homogeneous classroom, and instructed with like peers for part of the day. The study is intended to show which type of grouping is most conducive to student achievement in reading. The study provides educators with knowledge of the achievement results expected based on the way students are grouped for reading instruction. Results obtained from this study may be used to assist school leaders in determining instructional best practices for the teaching of reading having the greatest potential to produce positive student achievement outcomes.

Research Question

For this study, the researcher is seeking to answer the follow question: What are the effects of within-class homogeneous grouping versus the effects of between-class homogeneous grouping on student achievement in reading?

Research Hypotheses

H_{o1}. There is no significant difference in student achievement in reading by grouping form

Ho_2. There is no significant difference in the mean first grade DIBELS scores by grouping format.

Ho_3. There is no significant difference in the mean first grade STAR scores by grouping format.

Ho_4. There is no significant difference in the mean second grade DIBELS oral reading fluency scores by grouping format.

Ho_5. There is no significant difference in the mean second grade STAR scores by grouping format.

Ho_6. There is no significant difference in the mean third grade DIBELS oral reading fluency scores by grouping format.

Ho_7. There is no significant difference in the mean third grade STAR scores by grouping format

Ho_8. There is no significant difference in mean third grade MCT2 language-arts scores by grouping format.

Ho_9. There is no significant effect on student achievement in reading over a three-year period by grouping format.

Methods Overview

This study is conducted using a quasi-experimental, ex-post facto design and quantitative analysis of longitudinal data. The study focuses on determining the relationship between student achievement and two homogeneous grouping formats, within-class grouping and between-class grouping, as indicated by DIBELS oral reading fluency (DIBELS-ORF) subtest scores and STAR scores for students in grades one through three as well as MCT2 scores for students in grade three. Quantitative data is analyzed using a multivariate analysis of variance (MANOVA). According to Pearson Higher Education (n.d.), the MANOVA is used to determine how one independent variable with two levels impacts multiple dependent variables. In this study the independent variable is the grouping format. The independent variable has two levels: within-class grouping and between-class grouping, thereby indicating the first criteria for a MANOVA is met. Further indicating a MANOVA is appropriate, this study contains multiple dependent variables. The dependent variables are as follows: STAR scores, DIBELS-ORF Scores, and MCT2 scores. Analyzing multiple dependent variables and how they are influenced by the independent variable of grouping format is an improvement over previous studies in this field.

Limitations

This study is limited in several ways. For example, because the study involves the analysis of reading achievement based in part on the now obsolete Mississippi Curriculum Test, 2nd Edition, some of the results may not be applicable to expected results on current assessments. According to Pearson (2008), the MCT2 was designed to assess student achievement with regard to the Mississippi Curriculum Standards developed by the Mississippi Department of Education in 2007. Beginning with the 2013-2014 school year, the Mississippi Curriculum Standards were replaced with the Common Core State Standards. Then in 2015, these standards became known in Mississippi as the Mississippi College and Career Ready Standards. The MCT2 was given for the last time during the 2013-2014, school year. Another limitation is the subjects in this study were not randomly assigned to a treatment group. The subjects were assigned to groups in the past and assessments were completed at that time. The student data that is analyzed is quasi-experimental and ex-post facto in nature. Therefore, an experimental replication of the study may not reveal parallel results. One final limitation is teacher instructional processes or differences in instruction beyond the grouping format are not examined as a part of the study and cannot be controlled. Thus, the study does not take into account the years of teaching experience of the teachers involved, the educational background of the teachers, or specific abilities or inadequacies related to each teacher's use of differentiated instructional processes.

Definition of Terms

For the Purpose of this study, the following terms are defined as follows:

Ability Grouping: Grouping based on student ability level as determined by a test of cognitive ability or Intelligence Quotient. The term is often used interchangeably with

tracking or homogeneous grouping (Kerckhoff, 1986) but should not be confused with the term skill level grouping.

Between-Class Grouping: A format of homogeneous grouping in which students assigned to a heterogeneous classroom are regrouped for a portion of the day based on skill level and taught in a homogeneous classroom (Kulik & Kulik, 1992).

DIBELS: Acronym used for the Dynamic Indicators of Basic Early Literacy Skills assessment developed at the University of Oregon. The DIBELS assessment is a short series of tests used to evaluate the basic component skills of reading including but not limited to oral reading fluency (University of Oregon, 2015).

Differentiated Instruction: A way in which the teacher varies instruction for individual students or small groups of students in order to create the best learning environment possible for student achievement

Heterogeneous Grouping: A differentiated teaching model in which students of various skill or ability levels are placed in groups and learn together.

Homogeneous Grouping: A differentiated teaching model in which students of the same skill or ability level are placed in groups and learn together.

Mississippi Curriculum Test, 2nd Edition (MCT2): A criterion-referenced test developed by the state of Mississippi and was used to assess student achievement in grades three through eight. The test examined language arts and math skills and was aligned with the curriculum frameworks dated 2006. The MCT2 was developed to comply with the standards of No Child Left Behind and provided standard scale scores and achievement levels for students taking the test and was the basis for each school and district's achievement classification (Pearson, 2008).

Skill Level Grouping: Grouping based on a student's academic skill in any given subject area. Achievement tests, curriculum based measures, or skill assessments are typically used to determine skill level in a particular academic area (Condron, 2008).

STAR: The Standardized Test for the Assessment of Reading developed by the Renaissance Company. The STAR is regarded as a widely used, highly reliable and valid measure of reading.

Within-Class Grouping: A format of homogeneous grouping taking place within a heterogeneous classroom and occurs after the student's instructional level has been determined. For reading, teachers group students in small homogeneous groups for on-level instruction, enrichment, or remediation (Rowan & Miracle, 1983; Slavin, 1987).

Organization of the Study

Beyond the information provided in Chapter 1 of this study, Chapter 2 presents a review of literature on the topics of differentiated instructional grouping practices and outcomes and the reliability and predictive validity of assessment data related to the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), the Standardized Test for the Assessment of Reading (STAR), and Mississippi Curriculum Test 2nd Edition (MCT2) assessments. Literature outlining student achievement with regard to these topics is highlighted. Chapter 3 presents the methodology used in this study and includes information on the population, data collection, and quantitative analysis procedures. Chapter 4 provides the results of the statistical analysis. Finally, Chapter 5 presents a summary of the results and a discussion concerning the research outcomes.

Summary

According to the National Center for Educational Statistics (n.d.), less than 40% of students in grades four and eight in the United States scored proficient or higher in the area of

reading as determined by the 2013 National Assessment of Educational Progress (NAEP). In order to combat this serious problem, reading experts suggest ineffective instructional practices must be reduced and best practices must be increased (Carbo, 2007) with regard to reading instructional methods. While differentiated instruction is often noted as a best practice (Institute of Education Sciences, 2009) and is accomplished by homogeneously grouping students, there is little research available to guide teachers and instructional leaders when choosing the best homogeneous grouping format for their students. This study is needed in order to aid in establishing a best practice. With a new guide for what works, schools leaders are better able to implement the reading instructional format with the greatest potential to maximize student achievement.

CHAPTER TWO

Review of the Literature

Based on a review of current literature, there is a substantial body of work surrounding the topic of differentiated instruction as a best practice model for the teaching and learning of reading. One of the primary types of differentiated instruction used in the teaching of reading is homogeneous grouping. Many researchers have investigated the idea grouping can have dramatic effects on student achievement (Abadzi, 1984; Chueng, & Rudowicz, 2003; Condron, 2008; Gentry & Owen, 1999, Hong & Hong, 2009; Kerckhoff, 1986; Kulik & Kulik, 1992; McCoach, O'Connell, & Levitt, 2006; Meijnen & Guldemond, 2002; Rowan & Miracle, 1983). At the same time, there is also a component of the literature focusing on reading achievement predicted by varied assessments. However, this research fails to ascertain which format of homogeneous grouping produces the highest achievement results based on the analysis of more than one assessment over time.

For the purposes of this literature review, the research is divided into two primary categories: The Effects of Grouping on Student Achievement and Assessments as Indicators of Student Achievement in Reading. The review of the Effects of Grouping on Student Achievement is included in the subsections: Defining Grouping Models and Instructional Formats; The Early Grouping Debate and Psycho-Social Constructs of Grouping; Continued Research Into Achievement and the Psychosocial Effects of Grouping; The Divergence Effect; Overall Achievement as the Primary Focus of Grouping; and The Prevalence of Reading Grouping. The review of Assessments as Indicators of Student Achievement in Reading include

specific information regarding the reliability and predictive validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), Standardized Test for the Assessment of Reading (STAR), and the Mississippi Curriculum Test, 2nd Edition (MCT2) are also examined.

The Effects of Grouping on Student Achievement

Defining Grouping Models and Instructional Formats. One of the greatest tasks in the research on differentiated grouping is establishing an understanding of the differences between grouping models as well as the differences between the instructional formats used in conjunction with these models. Recognizing this issue, Kulik and Kulik (1992) and Slavin (1987) published papers on grouping specifically naming and addressing various forms of grouping. The purpose of these articles was to describe the types of grouping most widely used in schools in the United States. Although, the papers were simple commentaries on grouping, they are two of the most commonly referenced articles in the literature on grouping. These researchers described two primary differentiated grouping models: heterogeneous grouping and homogeneous grouping. Heterogeneous grouping is the process by which students of various skill or ability levels are placed in groups and learn together. Homogenous grouping is the process by which students of the same skill or ability level are placed in groups and learn together.

Kulik and Kulik (1992) and Slavin (1987) also identified two basic instructional formats primarily used for implementing the homogeneous grouping model. These include within-class grouping and between-class grouping. Within-class grouping typically denotes an instructional format in which students in a heterogeneous classroom are placed in small homogeneous groups within that classroom and receive their instruction based on their skill level. This is the primary format used with leveled-reading programs. The aforementioned authors also discussed the

between-class grouping format. Between-class grouping typically denotes an instructional format in which students are placed in a homogeneous classroom based on ability or skill level as determined by an intelligence quotient (IQ) measure or skill assessment. As part of the between-class grouping format, the students receive all or at least a portion of their instructional day in a class with homogeneous peers. In the full day version students spend all day with homogeneous peers.

The Early Grouping Debate and Psychosocial Constructs of Grouping. Early investigations into homogeneous grouping focused on the central idea wherein dividing students into groups and labeling them according to ability or skill level was inherently negative (Abadzi, 1984; Rowan & Miracle, 1983). Fearing homogeneous grouping somehow diminishes self-esteem, peer relationships, and even teacher expectations, researchers have set out to analyze the construct of grouping.

In 1983, Rowan and Miracle examined the idea homogeneous grouping affects peer relationships and in turn, peer relationships affect student achievement—a hypothesis referred to as the differential peer process. Secondly, the researchers proposed teacher expectations differ based on student ability or skill level and these expectations affect student achievement—a hypothesis referred to as the differential instruction hypothesis. In order to test these hypotheses, the researchers collected grouping and achievement data on 148 students in a large southern school district. All students in the sample were fourth graders and had been assigned to reading classes based on skill level. Two grouping formats were used simultaneously. The students were first grouped in a between-class format by skill level then grouped further within the classroom. Groups were formed based on student achievement scores on the Iowa Test of Basic Skills (ITBS). Rowan and Miracle (1983) observed the classroom samples and gathered

data on the number of peer and teacher interactions observed. The data were treated as interval scores. Student achievement was determined by growth measures on the ITBS. Correlation and regression techniques were used to analyze the relationship between peer and teacher interactions and achievement outcomes.

The results of the study rejected the differential peer hypothesis and showed no statistically significant correlations between peer interactions and student achievement. Likewise, the researchers rejected their differential instruction hypothesis. No statistically significant correlation was found to exist in regard to the relationship between teacher interactions and student achievement. Taken as a whole however, homogeneous grouping was found to be predictive of achievement. Higher-level homogeneous groups made significant gains over similar skill level students in the heterogeneous setting, while student in the lower level groups learned at the same rate as similar skill level students in the heterogeneous setting.

Still suspecting grouping has inherently negative outcomes on the psyche of students, Abadzi (1984) also examined grouping in a sociological sense. The purpose of the study was to analyze the effects of homogeneous grouping on student achievement and self-esteem. The subjects of the study were approximately 600 students from eight schools located in a large Texas school district. The schools were randomly selected from a larger population of 21 schools in the district. All students in the district were placed in homogeneous groups in the fourth grade and were either considered to be regular or high level students. Groups were formed based on third grade achievement scores. For this study, Abadzi administered the Coopersmith Self-Esteem Inventory to determine the self-esteem of the students before and after grouping was implemented. Student scores on the Iowa Test of Basic Skills and the California Achievement Test were used to compare group achievement before and after grouping.

The results of the study suggested there were no significant achievement effects for students who were placed in homogeneous groups. Instead, Abadzi concluded the students performed similarly to the way they performed before being placed in homogeneous groups. Thus, high-level students scored high and regular level students scored in the average range. In terms of self-esteem, the study implied there were no significant differences in student self-esteem before homogeneous grouping was applied. However, after students were grouped for a year, their self-esteem scores widened. High-level students showed significantly higher self-esteem than regular level students indicating homogeneous grouping is a positive instructional format for students of a higher learning level.

Continued Research Into Achievement and the Psychosocial Effects of Grouping. After approximately thirty years of research into grouping and a general consensus into the idea grouping can positively influence student achievement, the debate over the use of homogeneous grouping has continued because researchers have harbored concerns over the psychosocial effects of grouping. For example, Meijnen and Guldemond (2002) analyzed the theory surrounding the reference processes of students grouped in homogeneous classrooms and heterogeneously grouped classrooms in order to determine how student learning and achievement is affected. The reference processes as they are called, examined how students view their academic level and how they view themselves in terms of personal self-concept. Two basic reference types are suggested—task oriented and social-emotional. Task-oriented students are those who like to work with others for the purpose of completing a task. Social-emotional oriented students are those who like to work with others because they like them.

To determine the effects of student reference processes and grouping on achievement, Meijnen and Guldemond (2002) studied 3,648 students in 176 elementary schools in the

Netherlands. Thirty classes were examined. Of the classes, 14 were heterogeneously grouped and 16 were homogeneously grouped. Information was ascertained for all students in regard to task-orientation reference, social-emotional reference, and academic achievement. The results of the study showed a significant correlation between heterogeneous grouping and the task-oriented reference. Students in heterogeneous groups enjoyed working with others to complete academic tasks more than their peers in homogeneous groups. However, the data suggested the task-oriented reference did not have a positive impact on achievement. Overall, the study indicated high-level students make greater gains in achievement when they are grouped homogeneously and low-level students maker greater gains when they are grouped heterogeneously.

Likewise, Cheung and Rudowicz (2003) furthered the research into the psychosocial impact of grouping. It was hypothesized that students who were placed in homogeneous groups for instruction would have lower self-esteem, lower academic self-concept, higher testing anxiety, and lower academic achievement than students who were placed in heterogeneous groups. It was further hypothesized students placed in higher-level groups would have higher self-esteem, higher academic self-concept, lower test anxiety, and higher academic achievement than their peers in lower-level groups.

To determine the benefits and problems associated with homogeneous grouping Cheung and Rudowicz (2003) surveyed 2,720 students who had initially been grouped by intelligence quotient (IQ) scores and their teachers in 79 junior high schools in Hong Kong. A questionnaire was given to the students in order to determine self-esteem, self-concept, and test anxiety. The students' final course grades were used as the measure of student achievement and teachers were asked to provide these grades. Regression analysis was used to determine the predictive nature of one variable on another. The results of the study showed homogenous grouping had a slight

negative effect on self-esteem and test anxiety and a slight positive effect on academic self-concept. However, the results were not statistically significant. No differences were noted for student achievement based on homogenous grouping. Thus, the first research hypothesis was not supported on any level. The research results also indicated there were no significant differences in the self-esteem, self-concept, and test anxiety between students in various ability groups. However, the results suggested students in higher-level groups made achievement gains at a greater rate than students in other level groups.

The Divergence Effect. Along with the research seeking to examine the psychosocial constructs of grouping, another primary focus of research has been to examine the problematic effects of grouping on student achievement or what Kerckhoff (1986) coined as the "divergence effect." Primarily the divergence effect theory was born out of the studies frequently showing gaps in achievement between low-level learners who were grouped homogeneously and high-level learners who were grouped homogeneously. Specifically, the theory surmises when grouped homogeneously, high-level students are likely to make gains over their heterogeneously grouped peers, but low-level students are not. Thus, the disparity between low-level learners and high level learners increases when homogenous grouping is applied.

In an early research effort to examine the divergence theory, Kerckhoff (1986) analyzed the effect of homogeneous grouping on secondary students in Great Britain. Kerckhoff addressed divergence, but hypothesized the effect is due to pre-existing factors. Although Kerckhoff's aim was to gain support for the opposing traditional hypothesis contending all students will make gains when grouped homogeneously, the results of the study supported the divergence effect. Using the National Child Development Study (NCDS) a longitudinal study conducted by the National Children's Bureau of London, Kerckhoff (1986) studied a sample of

over 11,000 students. The students were given achievement tests year after year of primary school before the homogenous grouping model began during their high school years. Therefore, a pattern of achievement could be derived pre and post grouping and the divergence effect could be controlled. A regression model was used to analyze the effects of homogeneous grouping. According to the results of Kerckhoff's (1986) study, homogeneous grouping does have a significant impact on student achievement. This applies to achievement in both mathematics and reading. When controlling for the divergence effect, the results still showed there is a significant discrepancy between the rates of achievement made by those in high-level groups and those in low-level groups. Thus, indicating the divergence effect is probable.

Over twenty years later, Condron (2008) also explored the divergence effect. Condron analyzed data collected for the Early Childhood Longitudinal Study-Kindergarten Cohort conducted by the United States Department of Education beginning in 1998 and published by the National Center for Education Statistics in 2002. The data included information collected on an initial kindergarten cohort of over 20,000 students from 100 sample locations throughout the United States. Expressly, Condron examined reading test scores for over 13,000 students in the first and third grades. The data included information on student skill level, socioeconomic factors, and the type of instructional grouping format used to teach reading. Students in the study were either grouped homogeneously using a within-class instructional format or were instructed in a heterogeneous classroom with no grouping applied. The students who were grouped homogeneously were divided into three independent categories identified as high, middle, and low. Based on the socioeconomic identifiers of the typical student in each group and the likelihood of being placed in a particular group, the students in the heterogeneous category were matched and compared to homogeneously grouped students. Condron conducted a series

of t-tests to determine the mean differences in reading skill scores between students who were homogeneously grouped in each of the three levels of instruction and their heterogeneously grouped matched peers. A hierarchical regression model also was used to predict reading gains for the groups.

The results of Condron's (2008) analysis indicated homogeneous within-class grouping of high-level students does have significant statistical impact on reading gains. Students who were placed in the low-level groups did not gain as many reading skills as their matched peers in heterogeneous classrooms. Conversely, the students who were placed in high-level groups gained significantly more skills than similar students who were instructed in ungrouped heterogeneous classrooms for reading. The study also showed there were no significant differences between students identified as middle level learners. Overall, Condron concluded homogeneous within-class grouping applied to reading instruction produces unequal skill and achievement gains for students, again supporting the theory of the divergence effect.

Overall Achievement as the Primary Focus of Grouping. Although concerns over the psycho-social effects of grouping and the idea that grouping will further separate the low achieving student from the high achieving student through the divergence effect have remained prominent in the research on grouping, other studies have focused directly on the achievement effects of homogeneous grouping. In one such study, Gentry and Owen (1999) examined the effects of the cluster grouping of high-level students versus no cluster grouping on student achievement. Cluster-grouping was defined as the placement of three to ten students of a given ability or skill level into a classroom of otherwise heterogeneous students—homogeneous within-class grouping. The researchers wanted to address the relationship of cluster grouping and

teacher perceptions, the connection between homogenous grouping and student learning, and school and classroom factors affecting student achievement.

The study addressed the questions of teacher perceptions and student achievement through the use of quantitative analysis. The question of school factors affecting cluster grouping and student achievement were determined though quantitative methods. Two schools were studied over a four-year period. One school included the cluster-grouping model; the other did not. The student sample size was approximately 165. Fourteen teachers and three administrators were interviewed. Achievement from the Iowa Test of Basic Skills and the California Achievement Test were used for comparison.

The results of the study showed teachers perceived the cluster-grouping model to be positive and teachers generally believed in their students' ability to be successful regardless of their skill level. Also, the quantitative results showed after three years of cluster-grouping, the treatment school students made greater gains than the students in the heterogeneously grouped school. Finally, the qualitative analysis of the school programs indicated the teachers in the cluster-grouped schools were positive about their students' ability to learn, used varied teaching methods, and grouped students within the classroom using multiple formats and methods. The administrators in the cluster-grouped schools were in support of the grouping process and believed it was an effective method and influenced student achievement.

Similarly, other studies have found homogeneous grouping to positively influence student achievement. McCoach, O'Connell, and Levitt (2006) examined the ways in which reading growth and achievement in kindergarten are affected by the use of homogeneous within-class grouping. It was hypothesized there would be a positive correlation between reading achievement and the amount of time spent in the within-class grouping setting. A sample of

10,191 kindergarten students was examined. The sample was taken from data collected for the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K) conducted by the United States Department of Education. Student reading scores provided in the ECLS-K study were based on print-word recognition, sound identification, word reading ability, vocabulary, and comprehension skills. The researchers compared fall and spring reading scores for the sample and determined a measure of growth. Teaching practices were examined to determine the frequency of within-class grouping experienced by the students. The correlation between time spent in a skill level group and student reading achievement was then calculated.

The results of the study suggested 70% of kindergarten teachers use homogeneous within-class grouping. However, the frequency of the grouping on average is limited to once a week and for not more than 30 minutes. Nevertheless, correlation data indicated the use of within-class grouping is positively related to student reading achievement in kindergarten (McCoach, O'Connell & Levitt, 2006).

Further studies into the specific area of grouping and reading achievement have also indicated grouping has a positive influence on student achievement. Specifically, Hong and Hong (2009) explored the idea the amount of time spent in a homogeneous grouping setting is indicative of student achievement and growth in reading. The researchers posed a question regarding how the ineffectiveness of homogeneous grouping may be related to too little time being spent in direct reading instruction. For the purpose of analysis, time/grouping categories were classified and denoted as high intensity within-class grouping, low-intensity within-class grouping, and no within-class grouping.

Data was therefore taken from the U.S Department of Education's Early Childhood Longitudinal Study-Kindergarten Cohort and set up to be evaluated using a two-way analysis of

variance. Grouping effects and time effects were compared to student achievement scores. Before this analysis could be made, Hong and Hong (2009) applied a highly complex model of compiling the data. The results of the study suggested the effects of grouping are related to the amount of time spent in grouped instruction. Homogeneous grouping appeared to be indicative of higher levels of reading growth only when there is a high level of time provided for reading. The key amount of time was suggested to be one hour per day (Hong & Hong, 2009).

The Prevalence of Reading Grouping. In spite of the fears concerning psycho-social problems and the mixed results showing homogeneous grouping to be more effective with some groups than others, homogeneous grouping is thought of as a widely used instructional practice. Chorzempa and Graham (2006) sought to determine the prevalence of grouping and specifically focused on within-class grouping as the means of teaching reading in grades one through three. They also sought to determine the effectiveness of homogeneous grouping based on prevalence.

Chorzempa and Graham (2006) randomly selected 494 teachers from both public and private schools across the United States from a database of 1.6 million teachers. The teachers in the sample were asked to complete a comprehensive survey including information about their classroom teaching and grouping practices, the academic levels of their students, and their feelings and beliefs about grouping. Of the sample2 teachers responded to the survey. Survey responses were analyzed using factor analysis, correlation-regression, and analysis of variance techniques to determine relationships between specific variables.

The results of the study showed over 60% of the participants used homogeneous within-class grouping as a part of their instruction. The analysis of survey data also indicated the statistically significant reasons why teachers choose to use within-class grouping. These include possessing positive feelings about grouping, having less experience as a teacher, and teaching in

a rural as opposed to urban school. Lastly, the researchers found students in lower reading groups spend significantly less time reading silently and answering abstract comprehension questions, but spend significantly more time being read to, reading orally to their teacher, and answering literal comprehension questions than students in average and above average groups. Thus, there are significant differences in instruction dependent upon the level of the instructional group.

Assessments as Indicators of Student Achievement in Reading

Dynamic Indicator of Basic Early Literacy Skills (DIBELS). The University of Minnesota developed the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) during the 1970's and 1980's as a simple way to assess the skills needed for children to become readers (Dynamic Measurement Group, n.d.) According to the Dynamic Measurement Group, the University of Oregon then began researching the DIBELS assessments for validity and reliability. Since that time, research into the validity and reliability of the DIBELS assessments as a predictor of student reading ability has been extensive. For example, a search of the One Search Database yields over 3,300 scholarly articles related to the subject of DIBELS.

DIBELS measures five essential components of basic early literacy (Dynamic Measurement Group, n.d). These components include phonemic awareness, alphabetic principles and phonics, vocabulary, comprehension, and oral reading fluency. Although each of these essential skills has been widely studied, DIBELS oral reading fluency (DIBELS-ORF) has been cited for its predictive nature in regard to overall student achievement in reading. (Goffreda, Diperna, & Pedersen, 2009; Munger & Blachman 2013; Paleologos & Brabham, 2011).

According to the researchers at the University of Oregon (2015), the DIBELS-ORF assessment contains passages for each grade level. The passages are read aloud by the student

for one minute. If words are left out, misread, or there is more than a three-second hesitation those words are scored as an error. The number of words read correctly per minute is the oral reading fluency score.

The reliability of the DIBELS-ORF assessment has been reported to be strong. According to the researchers at the University of Oregon (2015), DIBELS-ORF test-retest reliability is .92 to .97 and the alternate form reliability ranges from .89 to .94. These figures are based on statistics documented in several separate studies compiled and reported by the University of Oregon. Likewise, according to the Mental Measurement Yearbook (Brunsman, 2003), the reliability of DIBELS ranges from .80 to .90.

Along with the reliability of the DIBELS assessments, the criterion related validity of DIBELS has also been examined. Researchers at The University of Oregon (2015) report the criterion related validity to range between .52 and .91, while the Mental Measurement Yearbook (Brunsman, 2003) reports the average predictive validity to be .66. This suggests DIBELS is a moderately strong predictor of achievement in reading.

Other independent researchers have also sought to determine the validity of DIBELS-ORF in regard to overall student achievement. For example, Goffreda, Diperna, and Pedersen (2009), studied the predictive validity of DIBELS compared to two state tests of academic achievement. DIBELS-ORF scores were compared to the test results of second and third grade students on the Terra Nova California Achievement Test (CAT) and the Pennsylvania System of School Assessment (PSSA). Based on the results, DIBELS-ORF was shown to be a significant predictor of future reading proficiency. Similarly, Munger and Blachman (2013) studied the relationship of DIBELS-ORF to the overall reading achievement results of third graders on the New York State English Arts (NYSELA) Test, the Wechsler Individual Achievement Test,

Second Edition (WIAT II), and the Group Reading Assessment and Diagnostic Evaluation (GRADE). The correlation coefficients for the test were .56, .66, and .72 respectively, thus affirming the positive predictive validity between DIBELS-ORF and student achievement in reading.

While DIBELS can be used to predict a student's performance on a state or national assessment of reading, there are other purposes for which the test can be used in the school setting. In 2009, Hoffman, Jenkins, and Dunlap examined educators' use of and perceptions about DIBELS. The results of the study showed educators often use DIBELS assessments for identifying students who were at-risk for reading deficits, to aid in the development of reading interventions for students in various tiered groups, and to progress monitor students receiving intervention.

Because of the simplicity in administering the DIBELS assessment and the strong reliability and validity associated with the test used for purposes such as determining student intervention needs, the DIBELS assessment is used on both a national and international basis in the educational setting. According to the University of Oregon (2015) over 28,000 schools worldwide have used the DIBELS data system, a service provided to school in addition to the assessments that help track and manage DIBELS data. Over one million unique users per month access the DIBELS data system, a statistic suggesting millions of students are being assessed with the instrument each year.

Standardized Test for the Assessment of Reading (STAR). According to Learning (2014), the Standardized Test for the Assessment of Reading (STAR) also known as STAR Reading was developed in 1998 as a test to provide teachers with an accurate and easy way to determine a student's reading level. The assessment was used in conjunction with the

Accelerated Reader Program, a computer based reading program designed to encourage and accelerate students' reading levels regardless of ability.

According to Renaissance Learning (2014a), STAR is both a criterion and norm referenced assessment. The test is criterion referenced because it measures student achievement in reading against criterion standards. The test is also norm referenced because it compares the achievement of each student to other students taking the test nationally. STAR can also be used to measure student progress in reading and can be used to show growth from one assessment to the next. A reading level or grade equivalency is provided each time a student is tested. The skills examined in each assessment include: foundational skills such as print concepts and phonological awareness; language skills such as vocabulary acquisition and use; literature skills such as understanding key ideas and details; and informational skills such as sequencing and determining cause and effect. The assessment is computer based, comprised of 34 items per assessment, and can be completed in approximately 15 minutes.

Although the STAR can be used for multiple purposes, one common practice in schools is to use the test for interim assessments or benchmarking student reading level throughout the school year (Renaissance Learning, 2014a). Results of the assessment are presented as both scaled scores and grade equivalent scores. Scaled scores (SS) range from 0 to 1400 and can be used to compare student progress over time. Grade equivalency (GE) scores range from 0.0 to 12.9 representing levels from kindergarten to the end of twelfth grade. The GE scores are norm referenced and compare students taking the test to other students nationally. Thus, a student who scores a 1.5 is reading at a rate equivalent to the average first grader in the fifth month of school (Renaissance Learning, 2014a).

The Standardized Test for the Assessment of Reading is also a computer adaptive assessment. Therefore, based on a student's response to each question, the next item has a higher or lower level of difficulty (Renaissance Learning, 2014a). Computer adaptive assessments have been shown to have a statistically significant predictive validity with respect to student achievement (Clemens et, al., 2015).

According to Renaissance Learning (2014a), STAR has strong internal reliability and strong test-retest reliability. In a study of more the 1.2 million students who took the test between 2012 and 2013, the internal reliability was found to be .97. In the same study conducted by Renaissance, test-retest reliability was examined for 5,000 students per grade levels 1-12. For this group the test –retest reliability was calculated at .90.

Multiple studies have been conducted by Renaissance Learning (2014a) to examine the predictive validity of the STAR assessment. In 196 predictive validity studies involving nearly 1 million students, the overall correlations ranged from .68 to .86. The average correlation was a strong .81. In these studies STAR was shown to have significant statistical links to the ACT Explore and the state curriculum tests in 48 states across the nation. Included in this number is the Mississippi Curriculum Test 2nd, Edition (Renaissance Learning, 2014a).

Mississippi Curriculum Test, 2nd Edition (MCT2). The Mississippi Curriculum Test was designed to assess student achievement toward benchmark standards of the Mississippi Curriculum Frameworks. According to Pearson (2008), special studies were conducted by the Mississippi Department of Education in 2007 to ensure the alignment of the original Mississippi Curriculum Test (MCT) and the Mississippi Curriculum Frameworks. The results of the special study showed the need for a revision in the curriculum to meet depth of knowledge standards at various subject and grade levels. The revision in curriculum called for a revision in the state-

based test. Thus, the Mississippi Curriculum Test, 2nd Edition was developed and field-tested in 2007 and first administered as an operational test in the spring of 2008 (Pearson, 2008).

Specifically, the MCT2 Language Arts Assessment was a 60-item multiple- choice test, operationally administered from the spring of 2008 through the spring of 2013. The MCT2 Language Arts Assessment was comprised of questions related to the five strands of language arts including reading, writing, speaking, listening, and viewing. Student scores were reported in terms of scaled scores and proficiency levels minimal, basic, proficient, and advanced (Pearson, 2008).

A study of the Mississippi Curriculum Test, 2nd Edition Language Arts Assessment conducted by Pearson (2008), determined the internal test reliability to be .84. The criterion-related validity of the test was determined to range from .53 to .62. Specifically, the criterion-related validity of the 3rd grade assessment was .62 (Pearson, 2008).

The predictive validity of the MCT2 Language Arts assessment was also addressed. For example, Renaissance Learning (2010) examined the link between STAR reading assessment scores and MCT2 Language Arts scores for approximately 31,000 students in grades three through eight. After scaled score comparisons, the correlation coefficients ranged from .68 to .74. Based on the results of the study, Renaissance Learning developed an equivalency table of MCT2 scaled scores and corresponding STAR scaled scores needed to achieve the Mississippi proficiency levels of basic, proficient, or advanced. While the study was conducted to show STAR scores could predict a student's score on the MCT2, the converse could also be noted (Renaissance Learning, 2010).

Conclusion

It is evident there is substantial research in the areas of grouping and the effectiveness of

instructional grouping formats. However, grouping studies conducted in the past tend to have been done on a relatively small scale and establish their connection to student achievement through the use of one measure. It is also evident based on the review of literature that there is no recent thorough comparison of differentiated grouping formats in which student achievement in reading has been has been addressed in a longitudinal manner, thereby evidencing a void in the research. The proposed study suggested in the forthcoming methodology has the potential to fill this void and to address the interaction of these important aspects of differentiated instruction and reading achievement.

CHAPTER THREE

Methodology

Introduction

The following chapter describes the methodology of a quantitative study concerning reading achievement. The effects of differentiated instruction on student achievement in reading are investigated. Specifically, the study seeks to determine the effects of within-class homogeneous grouping versus the effects of between-class homogeneous grouping on student achievement in reading. A quasi-experimental, ex-post facto, longitudinal design is implemented.

The independent variable in this study is homogeneous grouping and is divided into two levels: within-class grouping and between-class grouping. The study includes test scores from students in grades one, two, and three. The dependent variables are DIBELS-ORF scores for students in grades one through three, STAR reading scores for students in grades one through three, and MCT2 Language Arts scores for students in grade three. MCT2 is not given to students in grades one or two.

Study Design

This study is designed to answer the research question: What are the effects of within-class homogeneous grouping versus the effects of between-class homogeneous grouping on student achievement in reading. These effects are determined quantitatively using a quasi-experimental, ex-post facto research design (Creswell, 2009). The study design allows for the collection and analysis of longitudinal data and is appropriate because valuable information concerning widely used instructional grouping formats and student achievement results are

yielded. Overall, by conducting the study, school administrators and other instructional leaders are provided with specific knowledge of differentiated grouping formats having the greatest impact on reading achievement scores as it applies to both national and state assessments. Therefore, best practices regarding the teaching of reading can be deduced.

Population

The population for this study comes from a school district in Mississippi. The total population consists of 1206 subjects who began first grade from the 2006-2007 school year through the 2009-2010 school year. The population is divided into an experimental group and a control group. The control group consists of students who began first grade in school years 2006-2007 and 2007-2008. The experimental group consists of students who began first grade school years 2008-2009 and 2009-2010. From these groups, a refined population sample of 240 was determined based on each subject's inclusion in the DIBELS and STAR assessments over a three-year period and in the MCT2 assessment during the third grade year. Specifically, all members of the population were placed in an excel file. Subjects who were missing scores for DIBELS ORF, STAR, or MCT were eliminated from consideration as a subject. Due to the complete lack of third grade DIBELS scores for students who began first grade in 2006-2007 no students beginning first grade during that year could be selected for inclusion in the sample. Likewise only 8 students who began first grade in the 2009-2010 school year were assessed using DIBELS during their third grade school year and are therefore underrepresented in the sample. Overall, the refined population sample yielded 120 subjects in the control group and 164 subjects in the experimental group. While a multivariate analysis of variance could be conducted regardless of the number of participants in each group, the outcomes are most significant when the number in each independent variable group is equal (Pearson Higher Education, n.d).

Therefore, using a random numbers generator, 44 students were eliminated from the experimental sample to match or equal the control group number. This sample yielded a 5.66% margin of error and a 95% confidence interval (Raosoft, 2004). The students were then assigned a number to represent the grouping format they took part in for instruction. The format is dependent upon the year they began first grade. Condition 1 was assigned to those who were grouped within the classroom or began first grade in the 2007-2008 school year. Condition 2 was assigned to the students who were grouped between classrooms or began first grade in the 2008-2009 or 2009-2010 school year.

Instruments

The first instrument used in this study is the Standardized Test for the Assessment of Reading (STAR). Specifically, STAR scaled scores are used. According to Renaissance Learning (2014b), STAR is a norm-referenced test, providing scores for one student compared to the scores of other students and it is a criterion-referenced test, providing scores based on standard criteria. The reliability of the test has been measured using split-half reliability and test re-test reliability. In a split-half reliability measure with the number of subjects equaling 818,064, the Spearman-Brown coefficient was calculated at .918 showing good reliability. Likewise, in a measure of test re-test reliability with 64,472 subjects, the Pearson correlation was calculated at .85 also showing good reliability. Furthermore, STAR has also been shown to be valid in a meta-analysis comparing STAR scores to multiple assessments. The STAR Reading Technical Manual (Renaissance, 2014b) states:

> Using 569 correlation coefficients, the overall estimate of the validity of STAR Reading is 0.78, with a standard error of 0.001. The 95 percent confidence interval allows one to conclude the true validity coefficient for STAR Reading is approximately 0.78. The

> probability of observing the 569 correlations reported in Tables 24–27 if the true validity were zero, would be virtually zero. Because the 569 correlations were obtained with widely different tests, and among students from twelve different grades, these results provide strong support for the validity of STAR Reading as a measure of reading skills. (p. 89).

The second instrument to be used in the study is the Dynamic Indicators of Basic Early Literacy Skills, Oral Reading Fluency (DIBELS-ORF) assessment. DIBELS-ORF is a test of accuracy and fluency when reading connected text (University of Oregon, 2015). Students are tested on a one-on-one basis with the teacher. DIBELS-ORF scores are reported in the number of words read correctly in one minute. These scores are then translated into a range, identifying students as being on track for reading success or in need of academic intervention in reading. Researchers at the University of Oregon further reported the DIBELS-ORF test-retest reliability to be .92 to .97, the alternate form reliability to range from .89 to .94, and the criterion related validity to range from .52 to .91 based on a meta-analysis of several studies. These statistics indicate moderate to high levels of overall reliability and validity for this subtest.

The final instrument to be used in the study is the Mississippi Curriculum Test, 2nd Edition (MCT2) language arts assessment for third grade. The term language-arts is synonymous with reading. The test has been deemed a valid and reliable measure of third grade reading achievement. For example, The Mississippi Curriculum Test, 2nd edition Technical Manual (Pearson, 2008) showed a Cronbach's Alpha reliability coefficient of .90 for the third grade test of language arts. Similarly, criterion validity measures for student classroom performance and scaled scores third grade language arts revealed a moderate to strong correlation of .65.

Procedures

Before the study began the researcher sought approval from her dissertation committee and the Institutional Review Board of the University of Mississippi to conduct the quasi-experimental study using archival data. Upon approval of the study, the researcher was given access to the student data requested from the Mississippi school district. The superintendent of the school district preapproved the use of this data by the researcher. The collected data did not contain student identifiers, thus making the data anonymous. Similar data has been provided by the school district to individual local educational researchers and a local university.

The data contained test scores from STAR, DIBELS-ORF, and MCT2 for students who began first grade in the years 2006-2007, 2007-2008, 2008-2009, and 2009-2010. These data sets were selected based on continuity of instructional formats used to teach reading and of the assessment data available for analysis. Table 1 depicts the data and shows the logical progression and structure for this study.

Table 1

Grouping Model, Assessments Given, and Cohort Structure for Grades 1-3

School Year Students Began First Grade	State Test Given in 3rd Grade	Reading Assessments Given	Grouping Format Used in First Grade	Grouping Format Used in Second Grade	Grouping Format Used in Third Grade
2006-2007	MCT2	DIBELS STAR	1st Grade (1)	2nd Grade (1)	3rd Grade (1)
2007-2008	MCT2	DIBELS STAR	1st Grade (1)	2nd Grade (1)	3rd Grade (1)
2008-2009	MCT2	DIBELS STAR	1st Grade (2)	2nd Grade (2)	3rd Grade (2)
2009-2010	MCT2	DIBELS STAR	1st Grade (2)	2nd Grade (2)	3rd Grade (2)
2010-2011	MCT2	DIBELS STAR	1st Grade (2)	2nd Grade (2)	3rd Grade (1)
2011-2012	MCT2	DIBELS STAR	1st Grade (2)	2nd Grade (1)	3rd Grade (1)
2012-2013	PARCC	DIBELS Next STAR	1st Grade (1)	2nd Grade (1)	3rd Grade (1)

Control Group/Within-class Grouping for Three Consecutive School Years: Condition 1

Experimental Group/Between-class Grouping for Three Consecutive School Years: Condition 2

Scores from a three-year period from first through third grade were placed in an excel file for each anonymous subject in the within-class grouping cohort and the between-class grouping cohort. Final or end of the year STAR scale scores and DIBELS-ORF scores for each subject for their first through third grade years were entered into the excel file. A beginning of the year STAR score for students was available for first grade students was also entered into the file to serve as baseline data for the repeated measures analysis. Finally, an MCT2 reading/language arts scores for the third grade year was entered into the excel file. A column of data was also entered for the independent variable of grouping format. Within-class grouping was noted as Condition 1 and between-class grouping was noted as Condition 2. Data was exported from the Excel file to SPSS (version 23) for statistical analysis.

Hypothesis

$H_{o1.}$ There is no significant difference in student achievement in reading by grouping format.

$H_{o2.}$ There is no significant difference in the mean first grade DIBELS scores by grouping format.

$H_{o3.}$ There is no significant difference in the mean first grade STAR scores by grouping format.

$H_{o4.}$ There is no significant difference in the mean second grade DIBELS oral reading fluency scores by grouping format.

$H_{o5.}$ There is no significant difference in the mean second grade STAR scores by grouping format.

$H_{o6.}$ There is no significant difference in the mean third grade DIBELS oral reading fluency scores by grouping format.

H_{o7}. There is no significant difference in the mean third grade STAR scores by grouping format

H_{o8}. There is no significant difference in mean third grade MCT2 language-arts scores by grouping format.

H_{o9}. There is no significant effect on student achievement in reading over a three-year period by grouping format.

Statistical Tests and Data Analysis

In this study, the dependent variables are the scores from the STAR reading and DIBELS (ORF) assessments for first through third grade and the MCT2 reading/language arts scores for third grade. A baseline score was taken at the beginning of first grade for both DIBELS and STAR to be used in the repeated measures analysis. The other scores reflect end-of-year scores and are used in the general multivariate analysis of variance. The independent variable is the grouping format used for instruction. The independent variable has two levels: (1) the within-class grouping format and (2) the between class grouping format.

The data is analyzed using SPSS (version 23). Assumption testing was completed prior to conducting the statistical tests. Assumption tests are described in detail in Chapter 4. A Multivariate Analysis of Variance (MANOVA) was then conducted to examine the multivariate effect overall and at each grade level. According to Pearson Higher Education (n.d.), the MANOVA is appropriate because there are multiple dependent variables and one independent variable with two levels. The MANOVA indicates the multivariate outcome and describes how the independent variable of grouping format impacts the combination of dependent variables. As a part of MANOVA the correlation between the dependent variables must be observed. If the correlation is too high, the multivariate outcome must be rejected and the univariate outcomes

need to be analyzed using multiple ANOVAS. Univariate outcomes indicate the effect of the independent variable on each dependent variable separately (Pearson Higher Education, n.d.).

SPSS (version 23) reports several multivariate outcomes. Each outcome class is based on the number of variables. For the structure of this study Pillai's Trace is used. Pillai's Trace is considered to be a powerful test and can be used with any number of independent variable groups. An important factor for Pillai's Trace is it is most powerful when sample sizes are equal. Thus, the researcher randomly matched the refined population sample for each level of the independent variable (Pearson Higher Education, n.d).

In an extension of the MANOVA, a repeated measures analysis was conducted. The repeated measures test takes into account the STAR and DIBELS-ORF tests over a three-year period from first grade to third grade. Repeated measures analysis is appropriate for use with longitudinal data, according to Pearson Higher Education (n.d.), and should be used when each subject in the study has been tested using the same measure over several periods in time. Each subject in this study was measured each year over a three-year period on DIBELS-ORF and STAR. The MCT2 is not a repeated measure and is not be a part of this portion of the data analysis. The repeated measures analysis shows how the independent variable impacts the dependent variables over time and therefore helps to determine if grouping format has an effect on student achievement over several years of implementation.

Sample Quantitative Data Structure

Scale Scores for DIBELS-ORF, STAR, and MCT2 Language Arts are recorded as numerical values and remain as such in the SPSS (version 23) data view table by grade level. A number was assigned to each grouping formats. Within-class grouping, a format in which students are placed in a heterogeneous classroom but were regrouped within that classroom for

some reading instruction equals condition one (1). Condition 1 is also be referred to as the control group. Between-class grouping, a format in which students were placed in a homogeneous classroom and received reading instruction within that classroom with same level peers equals condition two (2). Condition 2 is referred to as the experimental group. A One-Way MANOVA was then run to address the multivariate effect for the dependent variables on reading achievement. A between-subjects analysis then examined the effects of each dependent variable at each grade level. Finally, for the repeated measures portion of the analysis, a baseline data point was added to SPSS (version 23) for both the STAR and the DIBELS-ORF test. Table 2 indicates a sample SPSS (version 23) data input structure for a One-Way Multivariate Analysis of Variance (MANOVA) with repeated measures.

Table 2

Sample Input Data for SPSS for a Repeated Measures MANOVA

Subject/ Student	Scale Score for DIBELS Baseline (DV1x)	Scale Score for DIBELS-1st (DV1a)	Scale Score for DIBELS-2nd (DV1b)	Scale Score for DIBELS-3rd (DV1c)	Scale Score for STAR-Baseline (DV2x)	Scale Score for STAR-1st (DV2a)	Scale Score for STAR-2nd (DV2b)	Scale Score for STAR-3rd (DV2c)	IV Grouping Format Condition
1	8	15	36	46	120	230	300	410	1
2	22	38	56	120	138	261	343	586	2
3	89	112	125	123	132	532	615	694	1
4	92	122	127	133	123	478	576	643	2
5	45	68	78	103	165	305	369	427	1
6	32	57	67	89	209	489	562	995	1

Representative of 240 subjects

DV1x= Student Scale Score for DIBELS ORF, Beginning of First Grade

DV1a=Student Score for DIBELS ORF, End of First Grade

DV1b=Student Score for DIBELS ORF, End of Second Grade

DV1c=Student Score for DIBELS ORF, End of Third Grade

DV2x= Student Scale Score for STAR, Beginning of First Grade

DV2a=Student Scale Score for STAR, End of First Grade

DV2b=Student Scale Score for STAR, End of Second Grade

DV2c=Student Scale Score for STAR, End of Third Grade

DV3a=Student Scale Score for MCT2, End of Third Grade

IV=Grouping Format:

Condition 1, Within-class grouping

Condition 2, Between-class grouping

Conclusion

A longitudinal quasi-experimental study using quantitative methods is appropriate to determine the effects of differentiated grouping formats on student achievement scores in the area of reading. The procedures describe the quantitative process used to determine the effects of program formats on achievement. The initial data analysis also addresses and describes how the researcher sought both the independent and interaction effects of program formats on reading achievement through a multivariate analysis of variance with repeated measures.

The results of the study provide school districts across the state of Mississippi with a new body of data aiding in the process of developing and implementing reading programs. The results of this study also allow for the determination of differentiated grouping as a researched-based best practice. Because best practices are also required to be part of the instructional process, this research could support the use of grouping to increase student achievement.

CHAPTER FOUR

Results

Introduction

The purpose of this quasi-experimental, ex-post facto study was to determine through quantitative analysis of longitudinal data, the effects of homogeneous differentiation on reading achievement. The population for this study came from a single school district in Mississippi. The population consists of 1206 subjects who began first grade in the 2006-2007, 2007-2008, 2008-2009, and 2009-2010 school years. The first two school years of students form the control group, while the second two-year span of students form the experimental group. The control group was taught reading in a within-class grouping format. The experimental group was taught reading in a between-class grouping format. From these groups, a refined population sample of 284 was selected based on each subject's inclusion in the DIBELS and STAR assessments over a three-year period and in the MCT2 assessment during the third grade year. An additional 44 subjects were randomly eliminated from the study in order to meet the best practice criteria of having equal sample sizes for each independent variable group analyzed using a MANOVA (Pearson Higher Education, n.d.).

In order to determine the effects of within-class skill level grouping versus the effects of between-class skill level grouping on student achievement in reading, student scores on the Dynamic Indicators of Early Literacy Skills (DIBELS), the Standardized Test for the Assessment of Reading (STAR), and the Mississippi Curriculum Test 2nd Edition (MCT2) Language Arts subtest were analyzed using a MANOVA. The MANOVA was used to determine if there are

statistical differences between the combined achievement scores for reading based on grouping format. An overall multivariate analysis was conducted to determine the differences in achievement by grouping format over a three-year period combined and each grade level was analyzed separately for between-subjects effects. The MANOVA was also able to determine if any of the achievement scores as separate units were affected by grouping format. Furthermore, using a repeated measures analysis, the study sought to determine if there is statistical significance related to the number of school years an instructional grouping format is implemented. The repeated measures analysis only included scores from DIBELS-ORF and STAR and included the addition of a baseline data point required for statistical analysis. The baseline data point for each test was a beginning of first grade score. The aforementioned analyses are discussed herein.

MANOVA Requirements and Assumptions Testing

Before statistical analysis can take place, assumptions regarding the data must be verified. According to Laerd Statistics (2016) there are several basic requirements and assumptions critical to a valid outcome in a multivariate analysis of variance. These assumptions include the following:

1. There must be two or more continuous dependent variables.
2. There must be at least one independent variable with two or more categorical groups.
3. There must be independence of observations or individual subjects in each group.
4. The sample size must be adequate with more subjects in each group than variables.
5. There must be no univariate outliers for each dependent variable.
6. There must be multivariate normality for each of the independent variable groups.

7. Dependent variables should not be overly correlated and thus there should not be multicollinearity.
8. There should be a linear relationship between each pair of dependent variables
9. There should be no multivariate outliers when considering the dependent variables as a group.
10. There should be homogeneity of variance-covariance.

Before the multivariate analysis of variance was conducted each of these assumptions was investigated. Some assumptions were examined with general observation, while others were calculated and analyzed using SPSS Version 23.

Continuous Dependent Variables. MANOVA assumes there are two or more continuous dependent variables. In this study there were three main dependent variables. These included scores on the DIBELS-ORF assessment, STAR assessment, and Mississippi Curriculum Test 2nd Edition. Both DIBELS-ORF and STAR consisted of a baseline score and three additional measure over a three-year period.

Categorical Independent Variables. Based on Laerd Statistics (2016) there should be at least one categorical independent variable with at least two levels or independent groups. In this study there was one independent variable—grouping format. Grouping format was divided into the independent groups of within-class grouping and between-class grouping; thus satisfying this assumption.

Independence of Observations. This data set contains only independent observations. Each score or data point is attached to a single subject. Each subject was selected based on beginning first grade in a particular school year and acquiring DIBELS-ORF and STAR scores over a four-year period. In order to be selected, each subject was also required to have a 3rd

Grade MCT2 score. None of these scores had any bearing or influence on any of the other scores for the individual subject or the scores of other subjects.

Adequate Sample Size. In order for a multivariate analysis of variance to be viable, there must be an adequate sample size. According to Laerd Statistics (2016), the sample size is sufficient when the number of subjects is greater than the number of variables. For this study, the total number of variables for each grade level analysis is two for first and second grade and three for third grade. The total number of variables is seven for the general MANOVA with combined effects. There are eight variables for the repeated measures analysis. The total sample is 240. There are 120 subjects in each of two categorical subgroups of the independent variable, far exceeding the number of total variables. Therefore, the sample size is adequate and the assumption is met.

Lack of Univariate Outliers. Each group or category of the independent variable should be free of univariate outliers for each of the dependent variables being analyzed (Laerd Statistics, 2016). According to Hinkle, Wiersma, and Jurs (2003), outliers are extreme scores in a distribution which may have an effect on the outcome of the data analysis. In order to determine if there were univariate outliers associated with the within-class grouping format and the between-class grouping format the data was examined using boxplot analyses. The boxplot analyses were run in SPSS Version 23. Data from each dependent variable was entered for a separate analysis. The dependent variables were named as follows: DIBELS-Baseline, DIBELS-1st, DIBELS-2nd, DIBELS-3rd, STAR-Baseline, STAR-1st, STAR-2nd, STAR-3rd, and MCT2-3rd.

The boxplot generated for each dependent variable compared the data by the independent variable of grouping format. The boxplots revealed all univariate outliers for the within-class

grouping format and the between class grouping format. SPSS Version 23 depicts outliers in two ways. First, according to Laerd Statistics (2016) the outliers are represented with an open circle or a star to represent their distance from the edge of the box or how far a score is from the normal distribution of scores. An open circle indicates a score is 1.5 box-lengths from the edge of the box. These scores are considered to be general outliers. An asterisk, however, indicates a score is three box-lengths from the edge of the box and is therefore, an extreme outlier. Secondly, outliers are denoted with a number corresponding to the data line in the SPSS data view table (see Appendix A). These output indicators allow the researcher to further examine the outlier data and determine the course of action for transforming the data, removing the outlier, or leaving the outlier in place to remain as part of the statistical analysis.

For this analysis, univariate outliers were found for all dependent variables at the general level of 1.5 box-lengths from the edge of the box. Extreme outliers of 3 box-lengths from the edge of the box were found for the dependent variables of DIBELS-Baseline and STAR-Baseline. The SPSS Version 23 boxplot output for each dependent variable by grouping format is summarized and depicted in Table 3 and Table 4. General outliers are shown in Table 3. Extreme outliers are shown in Table 4.

Table 3

Summary of General Univariate Outliers 1.5 Box-lengths from the Edge of the Box

Dependent Variable	Within-Class Grouping Total Number of Univariate Outliers at 1.5 Box-lengths from the edge of the box	Within-Class Grouping Univariate Outliers by Subject Numbers	Between-Class Grouping Total Number of Univariate Outliers at 1.5 Box-lengths from the edge of the box	Between-Class Grouping Univariate Outliers by Subject Numbers
DIBELS-Baseline	2	1, 106	2	141, 191
DIBELS-1st	3	1, 4, 14	3	141, 191, 233
DIBELS-2nd	4	1, 4, 14, 22	4	141, 191, 209, 211
DIBELS-3rd	4	1, 4, 14, 106	6	137, 141, 209, 21, 215, 233
STAR-Baseline	7	2, 26, 40, 51, 60, 92, 107	7	121, 174, 187 203, 221, 229 238,
STAR-1st	4	1, 4, 22, 102	2	141, 237
STAR-2nd	1	14	3	136, 137, 141
STAR-3rd	4	4, 7, 14, 102	6	137, 141, 191, 209, 211, 237
MCT2-3rd	1	65	0	

Table 4

Summary of Extreme Univariate Outliers 3 Box-lengths from the Edge of the Box

Dependent Variable	Within-Class Grouping Total Number of Univariate Outliers at 3 Box-lengths from the edge of the box	Within-Class Grouping Univariate Outliers by Subject Numbers	Between-Class Grouping Total Number of Univariate Outliers at 3 Box-lengths from the edge of the box	Between-Class Grouping Univariate Outliers by Subject Numbers
DIBELS-Baseline	2	4 14	0	N/A
DIBELS-1st	0	N/A	0	N/A
DIBELS-2nd	0	N/A	0	N/A
DIBELS-3rd	0	N/A	0	N/A
STAR-Baseline	12	1, 4, 7, 14, 22, 34, 62, 96, 100 106, 111, 116	15	137, 141, 148, 149, 161, 191, 192, 194, 209, 213, 216, 219, 232, 233, 237
STAR-1st	0	N/A	0	N/A
STAR-2nd	0	N/A	0	N/A
STAR-3rd	0	N/A	0	N/A
MCT2-3rd	0	N/A	0	N/A

The boxplot analysis revealed 30 outliers 1.5 box-lengths from the edge of the box for the within-class grouping format. There were 33 outliers 1.5 box-lengths from the edge of the box for the between-class grouping format. Overall, each of the dependent variables on each of the grouping formats showed 1.5 box-length outliers, with the exception of MCT2-3rd for the between-class grouping format. The box plot analysis also revealed 14 extreme outliers 3 box-lengths from the edge of the box for the within-class grouping format and 15 extreme outliers for the between-class grouping format. Twenty-seven of the extreme outliers were associated with the STAR-Baseline dependent variable and two were associated with the DIBELS-Baseline dependent variable. Therefore, the univariate outlier assumption was violated. When this assumption is violated steps must be taken and decisions must be made in regard to continuing with a multivariate analysis of variance (Laerd Statistics, 2016). Based on recommended procedures, the data outliers were first evaluated to confirm the scores in SPSS Version 23 were correct. There were no mistakes in regard to this step. Secondly, a judgment had to be made with regard to the conditions of the test and whether the tests were administered correctly. Because of the limited number of outliers it is reasonable to believe the tests were administered correctly. Furthermore, administrators for the DIBELS test are trained with regard to test administration, the STAR test is computer based and doesn't lend itself to administrative error, and MCT2 also required administrator training and an element of test administration security and accuracy due to it being a statewide accountability assessment. Lastly, a determination had to be made in the context of the accuracy of individual scores. Based on the highly valid and reliable nature of each of the assessments as noted in the review of literature, there is no reason to believe the test scores were invalid. The subject numbers derived from the boxplot analysis data also showed the outlier subject numbers to be consistent in most cases over multiple-tests, over

multiple years. Thus, it was determined the outlier scores were likely accurate and should be kept rather than transformed or thrown out.

Multivariate Normality. Multivariate normality is another assumption of a multivariate analysis of variance. According to Laerd Statistics (2016), this is a difficult assumption to analyze because it would essentially require the dependent variables to be analyzed simultaneously as a group, a task not possible in SPSS. In order to test for multivariate normality, the normal distribution of each dependent variable is tested along all categorical groups of the dependent variable (Pearson Higher Education, n.d.). This process is conducted using the Shapiro-Wilk Test of Normality. In order for a set of data for a given variable to be considered normally distributed, the Shapiro-Wilk significance level should be greater than .05 ($p > .05$). Table 5 shows the Shapiro-Wilk results for all dependent variables on each of the grouping formats of the independent variable. Based on the Shapiro-Wilk analysis, all variables with the exception of MCT2-3rd violated the assumption of normality with a significance level less than .05 ($p < .05$). Based on these results a decision had to be made with regard to moving forward with the multivariate analysis of variance. Although the majority of the dependent variables were not normally distributed, the scores are believed to be valid. It would be an extreme measure to transform the data or change the scores (Pearson Higher Education, n.d) particularly since the MANOVA is a statistical process deemed relatively robust to violations of normality (Laerd, 2016). Thus, the decision was made to keep the data and move forward with the analysis.

Table 5

Shapiro-Wilk Summary of Normality

Dependent Variable	Grouping Format	Statistic	df	Sig.	Violation of Normality
DIBELS-Baseline	Within-Class Grouping	.819	120	.000	Yes
	Between-Class Grouping	.873	120	.000	Yes
DIBELS-1st	Within-Class Grouping	.906	120	.000	Yes
	Between-Class Grouping	.950	120	.000	Yes
DIBELS-2nd	Within-Class Grouping	.956	120	.001	Yes
	Between-Class Grouping	.976	120	.032	Yes
DIBELS-3rd	Within-Class Grouping	.977	120	.036	Yes
	Between-Class Grouping	.974	120	.019	Yes
STAR-Baseline	Within-Class Grouping	.558	120	.000	Yes
	Between-Class Grouping	.658	120	.000	Yes
STAR-1st	Within-Class Grouping	.938	120	.000	Yes
	Between-Class Grouping	.957	120	.001	Yes
STAR-2nd	Within-Class Grouping	.975	120	.027	Yes
	Between-Class Grouping	.965	120	.003	Yes
STAR-3rd	Within-Class Grouping	.952	120	.000	Yes
	Between-Class Grouping	.950	120	.000	Yes
MCT2-3rd	Within-Class Grouping	.990	120	.551	No
	Between-Class Grouping	.990	120	.496	No

Test is significant at .05 ($p < .05$) indicating a violation of normal distribution.

Multicollinearity. In a multivariate analysis of variance it is important for the dependent variables be correlated but not be too closely related. Specifically, moderate correlations to stronger correlations not exceeding .90 are desired for the MANOVA (Laerd Statistics, 2016). This assumption is referred to as the lack of multicollinearity. In order to test this assumption bivariate correlations were run in SPSS Version 23 and the resulting Pearson correlations were examined. The results of this analysis (see Appendix B) revealed moderate to strong correlations for all dependent variables ranging from .414 for STAR-baseline compared to MCT2-3rd to .888 for DIBELS-Baseline compared DIBELS-1st (r=.414 to .888, p=.000). Based on these results there are no violations related to multicollinearity with this data.

Linearity. In a multivariate analysis of variance linear relationships are required for each pair of dependent variables as they relate to all groups of the independent variable (Pearson Higher Education, n.d.). According to Laerd Statistics (2016), if the dependent variables are not linear, the MANOVA can be run, however, the power of the test is reduced. Linearity is determined in SPSS Version 23 through the creation and analysis of scatterplot matrices. Both the within-class grouping format and the between-class grouping format were examined. Each pair of dependent variables was compared. The scatterplot initially shows only the data points for each of the two variables being compared. However, SPSS Version 23 scatterplot matrices (see Appendix C) can also be shown with a line of best fit for each pair of variables. In general, linearity is present if the data points form a straight line or move along the continuum with the line of best fit. Data points far removed from the line are likely outliers and could distort the linearity (Laerd Statistics, 2016). This was not the case for this data set. The greatest lack of linearity was present for the variables of STAR-Baseline and MCT2-3rd for both independent groups. Nevertheless, this poses no problem because STAR-Baseline and MCT2-3rd are never

compared. STAR-Baseline is only part of the repeated measures (RM) analysis and MCT2-3rd is not part of the RM analysis. Overall, each set of dependent variables showed a linear relationship and are acceptable for the MANOVA.

Multivariate Outliers. Similar to univariate outliers, multivariate outliers must also be examined before conducting a multivariate analysis of variance. While univariate outliers represent an individual data point from each dependent variable, multivariate outliers represent all data points for one subject taken as whole (Laerd Statistics, 2016). In order to determine multivariate outliers, a regression procedure called the Mahalanobis distance must be run in SPSS Version 23. Based on recommendations by Laerd Statistics (2016), the data output for Mahalanobis distance is then used to run a chi-square analysis using the degrees of freedom equal to the number of dependent variables in a multivariate data set. The chi-square critical values are then compared for significance at the .001 alpha level. ($p < .001$). For the purpose of this study, the data had to be run twice because there are two separate multivariate data sets. The first set of data is used in the MANOVA analysis and contains the dependent variables DIBELS-1st, DIBELS-2nd, DIBELS-3rd, STAR-1st, STAR-2nd, STAR-3rd, and MCT2-3rd. The results of regression analysis for Mahalanobis distance for the MANOVA showed six multivariate outliers and are summarized by grouping format in Table 6.

Table 6

Summary of Mahalanobis Distance for Multivariate Outliers: MANOVA

Grouping Format	Subject ID Number	Mahalanobis Distance	Chi Square Value
Within-Class Grouping	4	29.18657	.00006*
Within-Class Grouping	14	29.71676	.00004*
Within-Class Grouping	102	31.47541	.00002*
Between-Class Grouping	136	23.32941	.00069*
Between-Class Grouping	137	41.03304	.00000*
Between-Class Grouping	191	73.11269	.00000*

* Value less than .001 is a multivariate outlier.

The second set of data is used in the Repeated Measures analysis and requires baseline data and scores repeated over time. Therefore, the dependent variables DIBELS-Baseline and STAR-Baseline were added to the aforementioned set and the variable of MCT2-3rd was eliminated from the set for this portion of the study. The Mahalanobis distance results for the Repeated Measures analysis showed eight multivariate outliers. Table 7 provides a summary of results by grouping format.

Table 7

Summary of Mahalanobis Distance for Multivariate Outliers: Repeated Measures

Grouping Format	Subject ID Number	Mahalanobis Distance	Chi Square Value
Within-Class Grouping	14	26.46664	.00087*
Within-Class Grouping	1	28.03686	.00047*
Within-Class Grouping	102	29.73384	.00024*
Within-Class Grouping	34	32.79709	.00007*
Within-Class Grouping	4	41.07098	.00000*
Between-Class Grouping	211	47.53359	.00000*
Between-Class Grouping	137	65.77205	.00000*
Between-Class Grouping	191	73.57935	.00000*

* Value less than .001 is a multivariate outlier.

Overall, the results of the Mahalanobis distance analysis showed outliers for the MANOVA data set and the Repeated Measures data set. Thus, the multivariate outlier assumption was violated. After careful inspection of the multivariate outliers, it was determined these subject numbers coincided with univariate outliers which were previously determined to be valid. Because the MANOVA is robust to multivariate outliers when the sample size is large, as in this study, the decision was made to proceed with the analysis (Laerd Statistics, 2016).

Homogeneity of Variance-Covariance. The final assumption test conducted is the homogeneity of variance-covariance matrices. This test determines if the variables being studied are the same or similar for each group of the independent variable. This assumption is tested using Box's M test of the equality of covariance matrices. In order to conduct this test the

multivariate analysis of variance must be run. This test is specific to the variables being compared. Thus, Box's M was examined for the MANOVA set of variables of DIBELS-1st, DIBELS-2nd, DIBELS-3rd, STAR-1st, STAR-2nd, STAR-3rd, and MCT2-3rd. The test was also run for the Repeated Measures data set containing the addition of DIBELS-Baseline and STAR-Baseline variables and removing the MCT2-3rd variable. The results of the analyses are shown in Table 8.

Table 8

Box's M Test of Equality of Covariance Matrices

Analysis	N		Box's M	Sig.
MANOVA	Within-Class	120		
	Grouping		46.855	.020
	Between-Class	120		
	Grouping			
Repeated Measures	Within-Class	120		
	Grouping		75.004	.000*
	Between-Class	120		
	Grouping			

* Box's M is significant at ($p < .01$)

In order for the assumption of homogeneity of variance-covariance to be met, the test must not be statistically significant ($p < .01$). The analyses revealed there was homogeneity of variance-covariance for the MANOVA data set as assessed by Box's M ($p=.02$) and the assumption for this group of data was met. However, there was not homogeneity of variance-covariance for the repeated measures data set by Box's M ($p=.00$). Nonetheless, because the

sample size is equal for both groups of the independent variable, proceeding with the analysis is not problematic and can be handled appropriately by using Pillai's Trace instead of Wilk's Lambda (Laerd Statistics, 2016). Pillai's Trace is a more powerful and robust test of multivariate analysis of variance (Pearson Higher Education, n.d).

Multivariate Analysis of Variance Results

After the MANOVA assumptions were tested, the data set was deemed to be acceptable for analysis. A multivariate analysis of variance was then run in SPSS Version 23 to answer the research question: What are the effects of within-class homogeneous grouping versus the effects of between-class homogeneous grouping on student achievement in reading? While Ho_9 could not be addressed with the MANOVA, the following hypotheses were addressed:

Ho_1. There is no significant difference in student achievement in reading by grouping format.

Ho_2. There is no significant difference in the mean first grade DIBELS scores by grouping format.

Ho_3. There is no significant difference in the mean first grade STAR scores by grouping format.

Ho_4. There is no significant difference in the mean second grade DIBELS oral reading fluency scores by grouping format.

Ho_5. There is no significant difference in the mean second grade STAR scores by grouping format.

Ho_6. There is no significant difference in the mean third grade DIBELS oral reading fluency scores by grouping format.

Ho_7. There is no significant difference in the mean third grade STAR scores by grouping format.

Ho_8. There is no significant difference in mean third grade MCT2 language-arts scores by grouping format.

Descriptive Statistics. The descriptive statistics for the data set showed the mean scores and the number of subjects in each grouping format. The total mean for both groups over each variable was also shown. Table 9 provides a summary of the descriptive statistics and includes the total mean difference in the dependent variables.

Table 9

Summary of Descriptive Statistics for MANOVA

Dependent Variable	Grouping Format	Mean	Mean Difference	N
DIBELS-1st	Within-Class Grouping	73.28		120
	Between-Class Grouping	85.99		120
	Total	79.64	12.71	240
DIBELS-2nd	Within-Class Grouping	110.61		120
	Between-Class Grouping	117.22		120
	Total	113.91	6.61	240
DIBELS-3rd	Within-Class Grouping	131.14		120
	Between-Class Grouping	133.23		120
	Total	132.19	2.09	240
STAR-1st	Within-Class Grouping	247.15		120
	Between-Class Grouping	281.65		120
	Total	264.40	34.50	240
STAR-2nd	Within-Class Grouping	381.37		120
	Between-Class Grouping	421.43		120
	Total	401.40	40.06	240
STAR-3rd	Within-Class Grouping	520.45		120
	Between-Class Grouping	553.05		120
	Total	536.75	32.60	240
MCT2-3rd	Within-Class Grouping	155.75		120
	Between-Class Grouping	157.78		120
	Total	156.77	2.03	240

Based on the descriptive statistics, the mean for the between-class grouping format was higher for each dependent variable. The difference in scores based on grouping format ranged from 2.03 points on MCT2-3rd to 40.06 points on STAR-2nd. Generally, the MANOVA descriptive statistics showed a regression toward the mean for the between-class grouping format by the end of third grade.

Multivariate Test. The primary results of the multivariate analysis of variance are shown in the form of four statistical test values (Laerd Statistics, 2016). These tests are Hotelling's Trace, Wilk's Lamda, Roy's Largest Root, and Pillai's Trace. Determining which statistical test should be used in a given analysis is based on the number of independent variable groups and the number of subjects in each sample. Because there are two independent variable groups or grouping formats and the sample number in each group is equal, Pillai's Trace was chosen. Pillai's Trace is considered a more powerful and robust test and better addresses violations of assumptions (Pearson Higher Education, n.d.; Laerd Statistics, 2016).

The results of Pillai's Trace indicated a significant statistical difference between grouping formats on the combined dependent variables, $F(7, 232)= 2.275$, ($p = .029$; $p < .05$). Therefore, H_{o1} is rejected. There is a significant difference in student achievement in reading by grouping format.

Test of Between Subjects Effects. The test of between subject effects is used to further analyze the data if the main MANOVA result is statistically significant. According to Laerd Statistics (2016), the between-subjects effects provides the one-way ANOVA results for each dependent variable and helps determine which of the variables were factors in the statistical significance of the MANOVA. A summary of the between-subjects effects for the dependent variables is shown in Table 10.

Table 10

Summary of Between-Subjects Effects for Grouping Format

Dependent Variable	Sig.
DIBELS-1st	.010*
DIBELS-2nd	.186
DIBELS-3rd	.637
STAR-1st	.039*
STAR-2nd	.027*
STAR-3rd	.143
MCT2-3rd	.132

* Significant when less than .05 ($p < .05$).

The results of the between-subjects effects for the dependent variables showed there is a statistical difference by grouping format for DIBELS-1st ($p = .010$, $p < .05$), STAR-1st ($p = .039$, $p < .05$), and STAR-2nd ($p = .027$, $p < .05$). Thus, H_{o2}, H_{o3}, and H_{o5} are rejected. There is a significant difference in mean 1st grade DIBELS oral reading fluency scores by grouping format. There is also a significant difference in mean 1st grade STAR reading scores by grouping format. Furthermore, there is a significant difference in mean 2nd grade STAR reading scores by grouping format. Finally, H_{o4}, H_{o6}, H_{o7}, and H_{o8} are supported and thus there are no statistically significant differences in DIBELS-2nd, DIBELS-3rd, STAR 3rd, and MCT2-3rd by grouping format.

Repeated Measures Analysis Results

A multivariate repeated measures (RM) analysis was run to further answer the research question: What are the effects of within-class homogeneous grouping versus the effects of between-class homogeneous grouping on student achievement in reading? H_{o9} was also addressed: There is no significant effect on student achievement in reading over a three-year period by grouping format. For this analysis the repeated measures were DIBELS-ORF and STAR reading. The tests were given as a baseline at the beginning of first grade. The scores were then measured in three equal intervals at the end of first grade, end of second grade, and end of third grade.

Descriptive Statistics for Repeated Measures Analysis. The descriptive statistics for the data set showed the mean scores and the number of subjects in each group of the dependent variable by grouping format. The total mean for both groups over each variable was also shown. Table 11 provides a summary of the descriptive statistics for the repeated measures analysis and includes the mean difference between the dependent variables. Based on the RM analysis, scores were higher in all instances for the between-class grouping format as shown previously with the MANOVA. Likewise, DIBELS-Baseline and STAR-Baseline were also higher for the between-class grouping format. Mean differences in scores increased from baseline to first grade on both DIBELS and STAR. However, by the end of third grade, the trend in scores showed a regression toward the mean.

Table 11

Summary of Descriptive Statistics for Repeated Measures Analysis

Dependent Variable	Grouping Format	Mean	Mean Difference	N
DIBELS-Baseline	Within-Class Grouping	43.73		120
	Between-Class Grouping	52.70		120
	Total	48.22	8.97	240
DIBELS-1st	Within-Class Grouping	73.28		120
	Between-Class Grouping	85.99		120
	Total	79.64	12.71	240
DIBELS-2nd	Within-Class Grouping	110.61		120
	Between-Class Grouping	117.22		120
	Total	113.91	6.61	240
DIBELS-3rd	Within-Class Grouping	131.14		120
	Between-Class Grouping	133.23		120
	Total	132.19	2.09	240
STAR-Baseline	Within-Class Grouping	105.98		120
	Between-Class Grouping	109.63		120
	Total	107.80	3.65	240
STAR-1st	Within-Class Grouping	247.15		120
	Between-Class Grouping	281.65		120
	Total	264.40	34.50	240
STAR-2nd	Within-Class Grouping	381.37		120
	Between-Class Grouping	421.43		120
	Total	401.40	40.06	240
STAR-3rd	Within-Class Grouping	520.45		120
	Between-Class Grouping	553.05		120
	Total	536.75	32.60	240

Multivariate Test for Repeated Measures. Pillai's Trace was the statistical test selected to interpret the repeated measures analysis. Pillai's Trace is the most powerful of the multivariate statistical tests and is robust to violations of multivariate assumptions when sample numbers among independent variable groups are equal as in this study (Laerd Statistics, 2016). The results of Pillai's Trace indicated there is no significant statistical difference between grouping formats on the combined dependent variables over test measures repeated three times after baseline, $F(3, 236)= 2.556$, ($p = .056$; $p > .05$). Therefore, H_{o9} is supported. There is no significant effect on student achievement in reading over a three-year period by grouping format. Because no significance was found through multivariate repeated measures analysis, SPSS Version 23 does not calculate between-subjects effects.

CHAPTER FIVE

Summary, Discussion, and Implications

Introduction

The purpose of this study was to address two forms of homogeneous differentiated instruction: within-class skill level grouping and between-class skill level grouping. The within-class grouping format is a structure in which students of the same skill-level are taught in small homogeneous groups within a heterogeneous classroom. The between-class grouping format is a structure in which students of the same skill-level are taken from a heterogeneous classroom, regrouped to homogeneous classrooms, and taught with like peers for part of the day. This quasi-experimental study was intended to determine which differentiated grouping format is most beneficial in terms of student achievement in reading.

Summary of Study

The population for this study came from a single, Mississippi school district. Second and third grade scores came from one elementary school, whereas the first grade scores came from a pre-k through first grade feeder school. The total population consists of 1206 subjects who began first grade beginning the 2006-2007 school year through the 2009-2010 school year. A refined population sample of 240 subjects took part in the study. The final population sample contained 120 subjects in the control group who were instructed in the within-class grouping format and 120 subjects in the experimental group who were instructed in the between-class grouping format.

The independent variable in the study was grouping format. The dependent variables in the study were DIBELS oral reading fluency scores, STAR reading scores, and MCT2 Reading Language Arts scores. A multivariate analysis of variance and a repeated measures analysis was used to answer the research question: What are the effects of within-class homogeneous grouping versus the effects of between-class homogeneous grouping on student achievement in reading? The following hypotheses were also addressed:

H_{o1} : There is no significant difference in student achievement in reading by grouping format.

H_{o2}: There is no significant difference in the mean first grade DIBELS scores by grouping format.

H_{o3}: There is no significant difference in the mean first grade STAR scores by grouping format.

H_{o4}: There is no significant difference in the mean second grade DIBELS oral reading fluency scores by grouping format.

H_{o5}: There is no significant difference in the mean second grade STAR scores by grouping format.

H_{o6}: There is no significant difference in the mean third grade DIBELS oral reading fluency scores by grouping format.

H_{o7}: There is no significant difference in the mean third grade STAR scores by grouping format.

H_{o8}: There is no significant difference in mean third grade MCT2 language-arts scores by grouping format.

H_{o9}: There is no significant effect on student achievement in reading over a three-year period by grouping format.

Summary of Results

The multivariate analysis took into account the dependent variables of DIBELS-1st, DIBELS-2nd, DIBELS-3rd, STAR-1st, STAR-2nd, STAR-3rd, and MCT2-3rd. The results of the multivariate analysis of variance indicated there is a significant statistical difference between grouping formats on the combined dependent variables ($p = .029$). H_{o1} was rejected and it was determined there is a significant statistical difference in student achievement in reading by grouping format. Data from the between-subjects analysis of the MANOVA indicated there were statistically significant differences by grouping format for the dependent variables of DIBELS-1st ($p = .010$), STAR-1st ($p = .039$), and STAR-2nd ($p = .027$). Therefore, H_{o2}, H_{o3}, and H_{o5} were rejected. While there was no statistical significant difference for the variables of DIBELS-2nd, STAR-3rd, MCT2-3rd these scores approached a statistical difference and showed a higher mean for the between-class grouping format.

In order to determine the relevance of grouping format over time, the repeated measures analysis took into account the dependent variables of DIBELS-Baseline, DIBELS-1st, DIBELS-2nd, DIBELS-3rd, STAR-Baseline, STAR-1st, STAR-2nd, and STAR-3rd. The results of the repeated measures analysis indicated there is no significant statistical difference between reading achievement based on grouping over a three-year period ($p = .056$). Therefore, H_{o9} was supported. However, it is critical to point out at $p = .056$, the data is critically close to a statistically significant difference supporting the between-class grouping format. A mere .002 change would result in a statistical significance.

Discussion of Findings

For this study the researcher wanted to analyze which differentiated grouping format is most conducive to reading achievement. The results showed there was an overall statistical

difference favoring the between-class grouping format. Thus, the study indicated increased achievement results could be expected by implementing the between-class grouping format for instruction in reading. However, the study also suggested the positive achievement results for between-class grouping may not be sustainable over time as indicated by the three-year repeated measures analysis. Based on these mixed results, several topics for discussion arose.

Exclusion of Data Points to Create Equal Groups. Although the multivariate analysis of variance procedures were implemented according to standard statistical guidelines recommended by Pearson Higher Education (n.d.) and Laerd Statistics (2016) the researcher was concerned over the omission of data points in the between-subjects grouping format for the purpose of matching the sample population numbers. In order to address this concern the researcher re-ran the MANOVA with the total population sample. The unequal sample numbers were 120 in the within-class grouping format and 164 in the between-class grouping format. Before the MANOVA was run, the testing assumptions were examined. All assumptions were parallel to those in the equal subjects study. Therefore, the data was deemed appropriate for MANOVA testing. The multivariate results indicated there was a significant statistical difference between student achievement scores in reading by grouping format. Pillai's Trace indicated the significance level to be .030 ($p < .05$). The calculation was nearly equal to statistical result for the present study in which sample sizes were equal (.029; $p < .05$).

Significance of Outliers. Another area of concern was the number of multivariate outliers present which might effect the overall outcome of the analysis. While it is within the procedures of MANOVA to keep the multivariate outliers, the researcher can also make a determination to remove the multivariate outliers (Laerd Statistics, 2016). Although the researcher in this case believed in the quality and accuracy of the data including the outliers, the

multivariate analysis of variance was re-run in order to support the idea the data was accurate and the overall outcomes were accurate. First, all assumptions testing was conducted. The results of the assumptions testing showed fewer univariate outliers and the same violations of normality. There were no violations of multicollinearity and the requirements for linearity and homogeneity of variance were met. The data set uncovered no further multivariate outliers. Therefore, there were no issues with running the MANOVA based on assumptions violations. The results of the MANOVA with the outliers removed indicated a Pillai's Trace multivariate statistic of .018 ($p < .05$). The calculation was similar to the statistical result for the MANOVA study result (.029; $p < .05$) and again showed there is a significantly statistical difference between student achievement scores in reading by grouping format with between-class grouping resulting in a higher mean multivariate score.

Further Repeated Measures Analysis. Due to the nature of the assumptions testing required for accurate investigation, further repeated measures analysis was beyond a reasonable scope of the present study. However, a separate and additional analysis could be noteworthy as the multivariate statistic is only .002 points away from statistical significance. If re-running the RM analysis produced similar findings as observed with the re-running of the MANOVA, statistically significant results might be found.

Implications for Further Research

The MANOVA analysis showed the greatest differences in overall student achievement by grouping format were present at the end of first grade. By third grade, the differences in overall achievement by grouping format had regressed. Similarly, the repeated measures analysis showed no significant statistical differences in reading achievement by grouping format over a three-year period. During their first grade year, the students in this study attended one

feeder school. From second grade to third grade, these students attended one receiving school. These findings lead to several implications and possibilities further research.

Administrator and Teacher Commitment. Based on the aforementioned results, it could be speculated there was a difference in the teacher and/or administrator commitment to the instructional grouping format due to the decreases in statistical significance when students moved from one school to another. Public records from the district's school board meetings, show that prior to implementing the between-class grouping format, the teachers and administrators of the feeder school proposed a change to the superintendent and school board. The proposal was presented by the school principal and signed by all teachers in first grade. No such proposal was recorded in future years for the receiving school.

Furthermore, the overall sample size for this study was only 240 even though there were 1206 subjects in the population. The researcher expected student attrition over the three-year period required for inclusion in the study, but at the same time, anticipated a sample of no less than 500. The reduction in sample size was largely based on the fact that a substantial amount of data was missing for grades two and/or three. For example, there were no second grade DIBELS scores for students in second grade during the 2007-2008 school year nor could any public record of the overall results could be found. While the specific reason for the lack of data is unknown, it is speculated the test was not given in 2007-2008 at the second grade level.

This aforementioned information taken as a whole implies there may have been a difference in the receiving school's evaluation processes concerning student achievement. Without commitment and buy-in from school staff, implementing and determining the success of instructional initiatives can be challenging and may not produce positive achievement results. Therefore, it would be worthwhile for a future researcher to examine the achievement results of

schools based on grouping formats selected with the addition of a survey analysis or qualitative study involving teachers and administrators who were involved in the processes. Research of this nature might explore the reasons why programming formats are chosen by administrators and how teacher and administrative commitment influence student achievement.

Instructional Scheduling and Academic Achievement. The current study did not take into account instructional scheduling and the time required for the instruction of reading and other academic subjects. Due to the nature of the within-class grouping format, students are divided into small groups within the classroom and instructed on their skill level. Essentially, the within-class grouping format requires more time for reading instruction, results in a decreased amount of time each student receives in on-level instruction, and likely decreases the amount of time remaining in the instructional day for the teaching of other subject areas. With the between-class grouping format, students are regrouped for a portion of the day with students on the same or similar skill-level. This grouping format requires less time the reading instructional period, provides the students with more time spent specifically on skill level, and potentially increases the amount of time left for instruction in other subject areas. With these ideas in mind, the current study could be expanded to explore the idea of instructional time management and the differences in time required for each grouping format. An expanded study might also address any correlations between the time spent in a specific grouping format and the second-hand effect on achievement in other subject areas. For example, achievement in mathematics could be examined for the present MANOVA study adding MCT2-Mathematics as an additional variable.

Long Term Outcomes for Differentiated Grouping Formats. Although there was no statistical significant difference over time between student achievement scores based on grouping format, the study only focused on the short-term span of three years. The question

remains if there are any lasting effects on student achievement based on grouping format. Specifically, does the grouping format used for reading instruction in the primary grades influence the overall K-12 educational outcome? Educational outcome is measured by college and career readiness. The current standard for measuring college and career readiness in Mississippi is the American College Test (ACT). ACT data will soon be available for the cohorts of students examined in this study. Therefore, the present study could be expanded to determine if there are any long-term differences between the ACT scores of students who were instructed in the within-class format versus those instructed in the between-class format. Although there are obviously many variables effecting student achievement over the course of one's school career, it would be noteworthy to determine if the statistical differences seen during the primary years of fundamental reading instruction correlated to a student's college and career ready outcomes.

Conclusion

The current study showed there was a statistical significant difference between reading achievement scores based on the differentiated grouping format implemented. Specifically, the study indicates between-class grouping produces higher mean achievement scores when compared to within-class grouping. However, there were no statistically significant differences between student achievement scores over a three-year period. Although there is a wealth of data supporting the implementation of differentiated instruction in schools, it is important for research efforts to continue to address the best practices in differentiated instructional processes as they are related to the essential skill of reading. By determining the best practices in reading instruction, school administrators will have valuable information which could assist them in making the best decisions regarding the planning and implementation of instruction. Likewise,

there appears to be a critical link between administrator and teacher commitment to the instructional program and student achievement outcomes. With best practices applied, and dedication to implementing the instructional program with fidelity, schools will have the greatest chance of improving student achievement.

REFERENCES

Abadzi, H. (1984). Ability grouping effects on academic achievement and self-esteem in a southwest school district. *Journal of Educational Research, 77*(5), 287-292.

Brunsman, B. A. (2003). Review of the DIBELS: Dynamic Indicators of Basic Early Literacy Skills (6th ed.). In B. S. Plake, J. C. Impara, & R. A. Spies (Eds.), The fifteen mental measurements yearbook (pp. 307-310). Lincoln, NE: Buros Institute of Mental Measurements.

Carbo, M. (2007). Best practices for achieving high, rapid reading gains. *Principal, 87*(2), 42-45. Retrieved from http://www.nrsi.com/docs/publications/pub_13_best_practices1.pdf

Chorzempa, B. F., & Graham, S. (2006). Primary-grade teachers' use of within-class ability grouping in reading. *Journal of Educational Research 98*(3), 529-541.

Chueng, C., & Rudowicz, E. (2003). Academic outcomes of ability grouping among junior high school students in Hong Kong. *The Journal of Educational Research, 96*(4), 241-254.

Clemens, N. H., Hagen-Burke, S., Wen, L., Cerda, C., Blakely, A., Frosch, J., . . . VanDerHayden, A. (2015). The predictive validity of a computer adaptive assessment of kindergarten and first-grade reading skills. *School Psychology Review 44*(1), 76-97.

Condron, D. J. (2008). An early start: Skill grouping and unequal reading gains in the elementary years. *The Sociological Quarterly* 49, 363-394.

Creswell, J. W. (2009). *Research design: Quantitative, qualitative, and mixed-methods designs.* (3rd ed.). Los Angeles, CA: Sage.

Daly, A. L., (2009). Rigid response in an age of accountability: The potential of leadership and trust. *Educational Administration Quarterly. 45*(2), 168-216.

Denton, C.A., Foorman, B. R., & Mathes, P.G. (2003). Perspective schools that beat the odds: Implications for reading instruction. *Remedial and Special Education, 24*(5), 258-261.

Dynamic Measurement Group (n.d.). What are dibels. Retrieved from https://dibels.org/dibels.html

Gentry, M., & Owen, S. V. (1999). The effects of whole school flexible cluster grouping on identification, achievement, and classroom practices. *Gifted Child Quarterly 43*(4), 224-241.

Goffreda, C. T., Diperna, J. C., & Pedersen, J. A. (2009). Preventive screening for early readers: Predictive validity of the dynamic indicators of basic early literacy skills. *Psychology in the Schools 46*(6), 539-552.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin Company.

Hoffman, A. R., Jenkins, J. E., & Dunlap, S. K. (2009). Using DIBELS: A survey of purposes and practices. *Reading Psychology* 30*,* 1-16.

Hong, G., & Hong, Y. (2009). Reading instruction time and homogeneous grouping in kindergarten: An application of marginal mean weighting through stratification. *Educational Evaluation and Policy Analysis 31*(1), 54-81.

Huebner, T. (2010). What research says about differentiated instruction. *Educational Leadership 67*(1), 79-81. Retrieved from http://www.ascd.org/publications/educational-leadership/feb10/vol67/num05/Differentiated-Learning.aspx

Institute of Education Sciences (2009). Using data to differentiate instruction. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/practice_guides/wwc_rrti_pg_rec02.pdf

Kerckhoff, A. C. (1986). Effects of ability grouping in British secondary schools. *American Sociological Review 44*(2), 298-338.

Kulik, J. A. & Kulik, C. L. C. (1992). Meta-analytic findings on grouping programs. *Gifted Child Quarterly 36*(2), 73-77.

Laerd Statistics. (2016). Statistical tutorials and software guides. Retrieved from https://statistics.laerd.com/

Massachusetts Model System for Teacher Evaluation (2012). Retrieved from http://www.doe.mass.edu/edeval/model/

McCoach, D. B., O'Connell, A. A., & Levitt, H., (2006). Ability grouping across kindergarten using an early childhood longitudinal study. *The Journal of Educational Research. 99*(6), 339-346.

Meijnen, G. W., & Guldemond, H. (2002). Grouping in primary schools and reference processes. *Educational Research and Evaluation, 8*(3), 229-248.

Mississippi Statewide Teacher Appraisal Rubric (2014). Retrieved from http://www.mde.k12.ms.us/docs/teacher-center/revised-m-star-rubric-june-2014.pdf?sfvrsn=2

Mississippi Curriculum Test, Second Edition Technical Manual for the 2007-2008 Administration. Retrieved from https://districtaccess.mde.k12.ms.us/studentassessment/DTC%20Resources/Program_Manuals/MCT2/MCT2%202007-2008%20Tech%20Manual.pdf

Munger, K. A., & Blachman, B. A. (2013). Taking a simple view of the dynamic indicators of basic early literacy skills as a predictor of multiple measures of third-grade reading comprehension. *Psychology in the Schools 50*(7), 722-737.

National Center for Education Statistics. (n.d.). Retrieved from http://www.nationsreportcard.gov/reading_math_2013/#/what-knowledge

No Child Left Behind Act of 2001, Pub. L. No. 107-110. Retrieved from http://www. ed.gov/ policy/elsec/leg/esea02/index.html

O'Connell, R. J., & White, G. P. (2005). Within the accountability era: Principals' instructional leadership behaviors and student achievement. *NASSP Bulletin 89*(645) 56-71.

Paleologos, T. M., & Brabham, E. G. (2011). The effectiveness of DIBELS oral reading fluency for predicting reading comprehension of high- and low-income students. *Reading Psychology 32*(1), 54-74. Retrieved from http://eric.ed.gov/?id=EJ9123

Pearson Higher Education. (n.d.). Multivariate analyses (14). Retrieved from http://www.pearson highered.com/assets/hip/gb/uploads/Mayers_IntroStatsSPSS_Ch14.pdf

Preckel, F., Gotz, T., & Frenzel, A. (2010). Ability grouping of gifted students: Effects on academic self-concept and boredom. *British Journal of Educational Psychology 80*(3), 451-472.

Raosoft. (2004). *Sample size calculator.* Retrieved from http://www. raosoft.com/samplesize. html

Reitzug, U. C., West, D. L., & Angel, R. (2008). Conceptualizing instructional leadership: The voices of principals. *Education and Urban Society 40*(6), 694-714.

Renaissance Learning. (2010). Relating star reading and star math to the Mississippi Curriculum Test, 2nd Edition MCT2. Retrieved from http://doc.renlearn.com/KMNet/R00451049GK E965.pdf

Renaissance Learning. (2014a). The research foundation for star assessments: The science of STAR. Retrieved from http://doc.renlearn.com/KMNet/R003957507GG2170.pdf

Renaissance Learning. (2014b). STAR reading technical manual. Retrieved from http://doc.renlearn.com/KMNet/R004384910GJF6AC.pdf

Rogers, K. B. (2007). Lessons learned about educating the gifted and talented: A synthesis of research on educational practice. *Gifted Child Quarterly 51*(4), 382-396.

Rowan, B., & Miracle, A.W. (1983). Systems of ability grouping and the stratification of achievement in elementary schools. *Sociology of Education.* 18, 133-134.

Slavin, R. (1987). Grouping for instruction in elementary school. *Educational Psychologist, 22*(2), 109-127.

Sparks, S. D. (2015). Differentiated instruction: A primer. *Education Week.* Retrieved from http://www.edweek.org/ew/articles/2015/01/28/differentiated-instruction-a-primer.html

Texas Teacher Evaluation and Support System (2014). Retrieved from http://tea.texas.gov/Texas_Educators/Educator_Evaluation_and_Support_System/Texas_Teacher_Evaluation_and_Support_System/

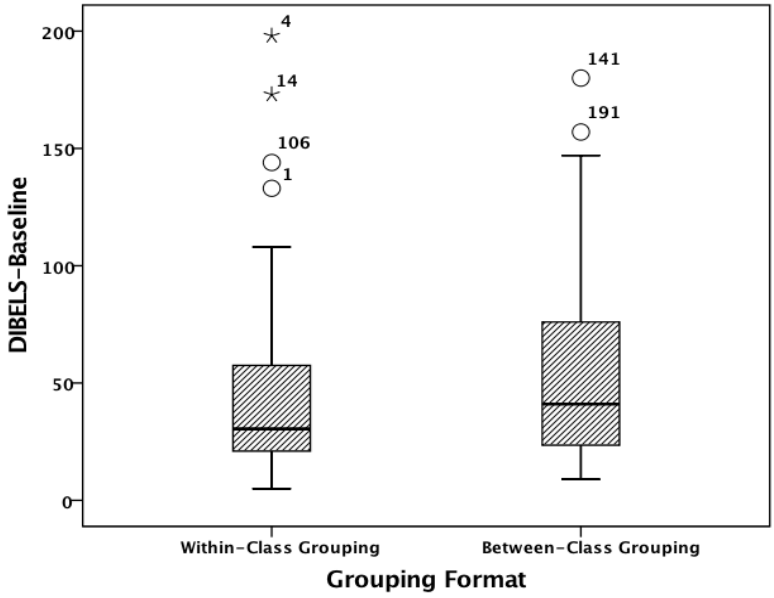
Tomlinson, C. A. (2000). Reconcilable differences standards-based teaching and differentiation. *Educational Leadership 58*(1), 6-11. Retrieved from http://www.ascd.org/ publications/ educational_leadership/sept00/vol58/num01/Reconcilable_Differences¢_Standards-Based_Teaching_and_Differentiation.aspx

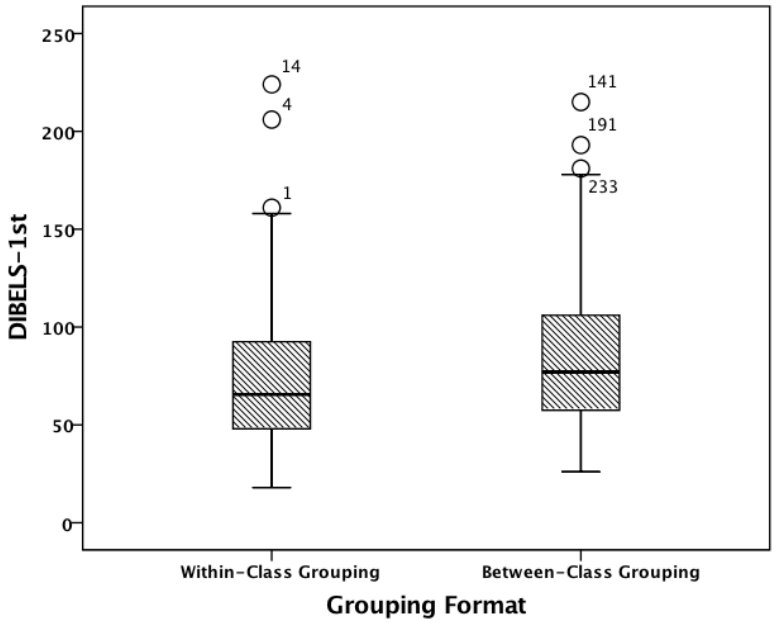
University of Oregon (2015). UO dibels data system. Retrieved from https://dibels.uoregon.edu/assessment/dibels/index

U. S. Department of Education. (n.d.). Laws and guidance/elementary and secondary education: ESEA flexibility. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility

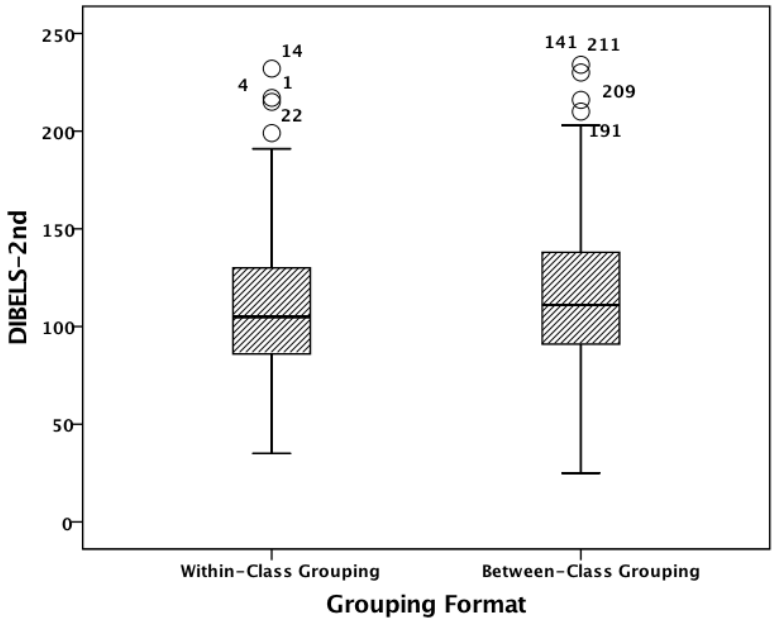
LIST OF APPENDICES

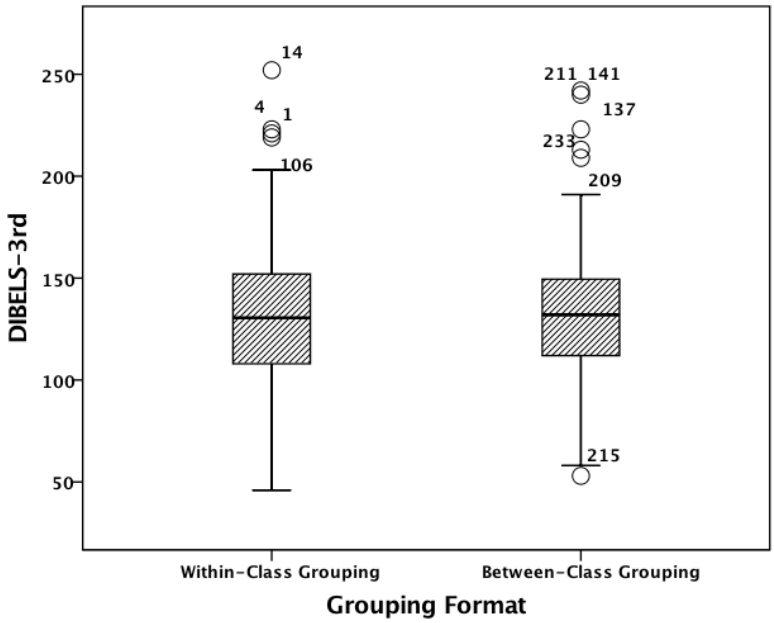
APPENDIX A: SPSS VERSION 23 BOXPLOT OUTPUT



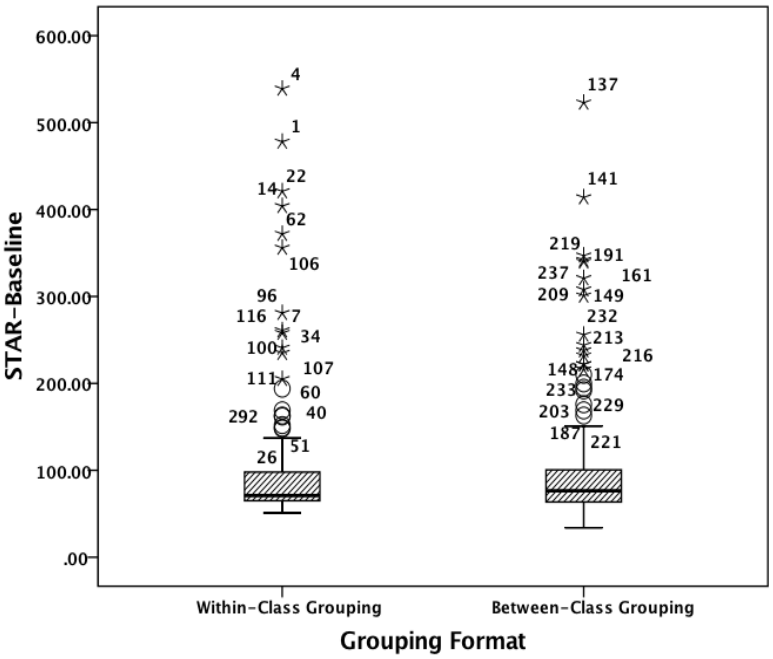


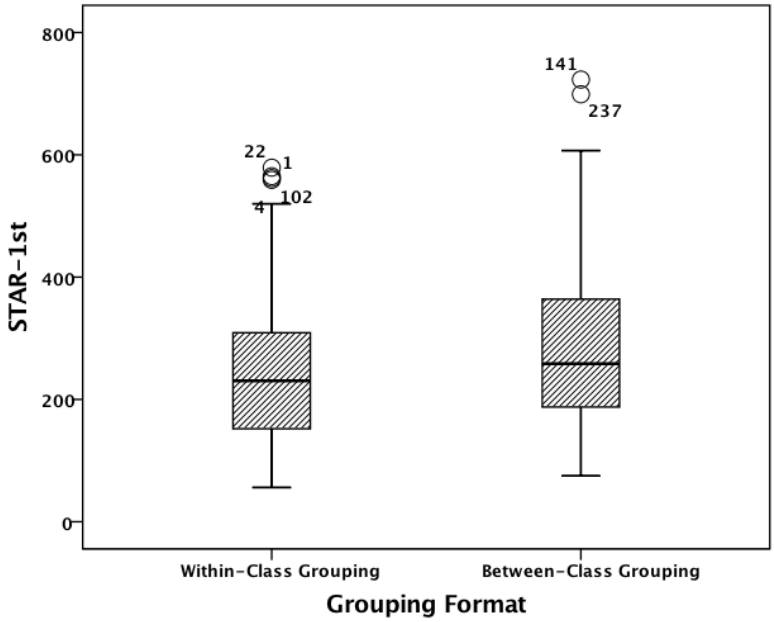
- o Univariate Outliers at 1.5 box-lengths from the edge of the box with numbers notating the subject number in SPSS Data View.
- * Univariate Outliers at 3 box-lengths from the edge of the box with numbers notating the subject number in SPSS Data View.





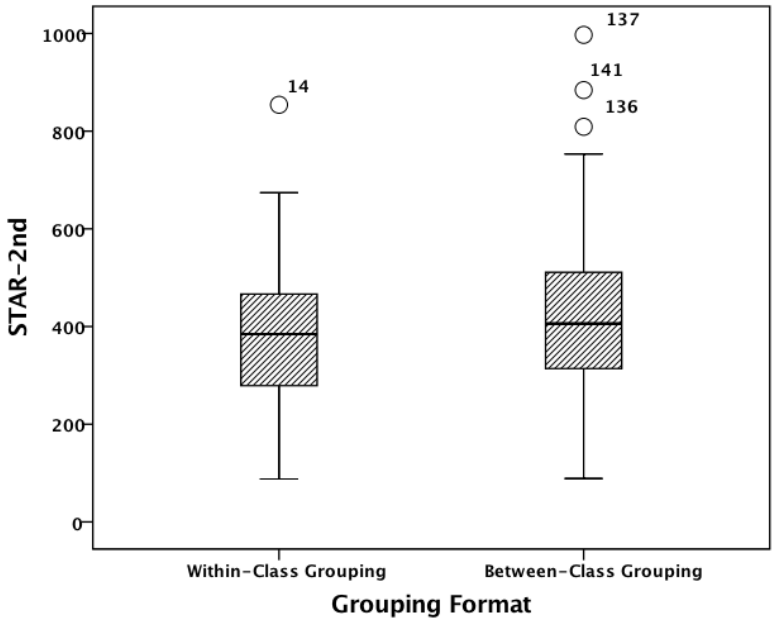
- Univariate Outliers at 1.5 box-lengths from the edge of the box with numbers notating the subject number in SPSS Data View.

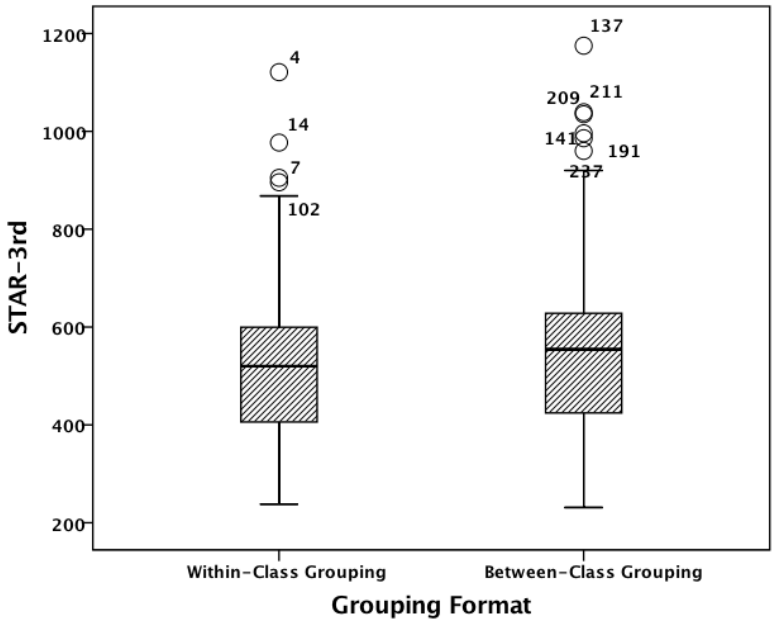




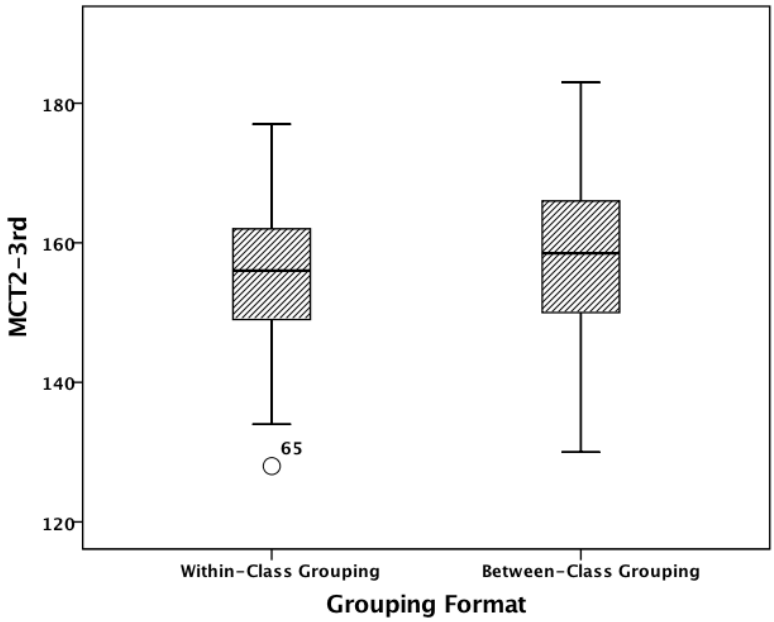
- Univariate Outliers at 1.5 box-lengths from the edge of the box with numbers notating the subject number in SPSS Data View.

* Univariate Outliers at 3 box-lengths from the edge of the box with numbers notating the subject number in SPSS Data View.





- Univariate Outliers at 1.5 box-lengths from the edge of the box with numbers notating the subject number in SPSS Data View.



- Univariate Outliers at 1.5 box-lengths from the edge of the box with numbers notating the subject number in SPSS Data View.

APPENDIX B: BIVARIATE CORRELATIONS FOR MULTICOLLINEARITY

Bivariate Pearson Correlations for Multicollinearity

		DV1x	DV1a	DV1b	DV1c	DV2x	DV2a	DV2b	DV2c	DV3a
DV1x	Pearson	1								
	Sig.									
DV1a	Pearson	.888**	1							
	Sig.	.000								
DV1b	Pearson	.725**	.866**	1						
	Sig.	.000	.000							
DV1c	Pearson	.642**	.774**	.834**	1					
	Sig.	.000	.000	.000						
DV2x	Pearson	.795**	.718**	.609**	.549**	1				
	Sig.	.000	.000	.000	.000					
DV2a	Pearson	.761**	.824**	.754**	.706**	.649**	1			
	Sig.	.000	.000	.000	.000	.000				
DV2b	Pearson	.629**	.712**	.686**	.710**	.550**	.787**	1		
	Sig.	.000	.000	.000	.000	.000	.000			
DV2c	Pearson	.621**	.699**	.709**	.733**	.563**	.746**	.824**	1	
	Sig.	.000	.000	.000	.000	.000	.000	.000		
DV3a	Pearson	.499**	.610**	.642**	.632**	.414**	.645**	.675**	.693**	1
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

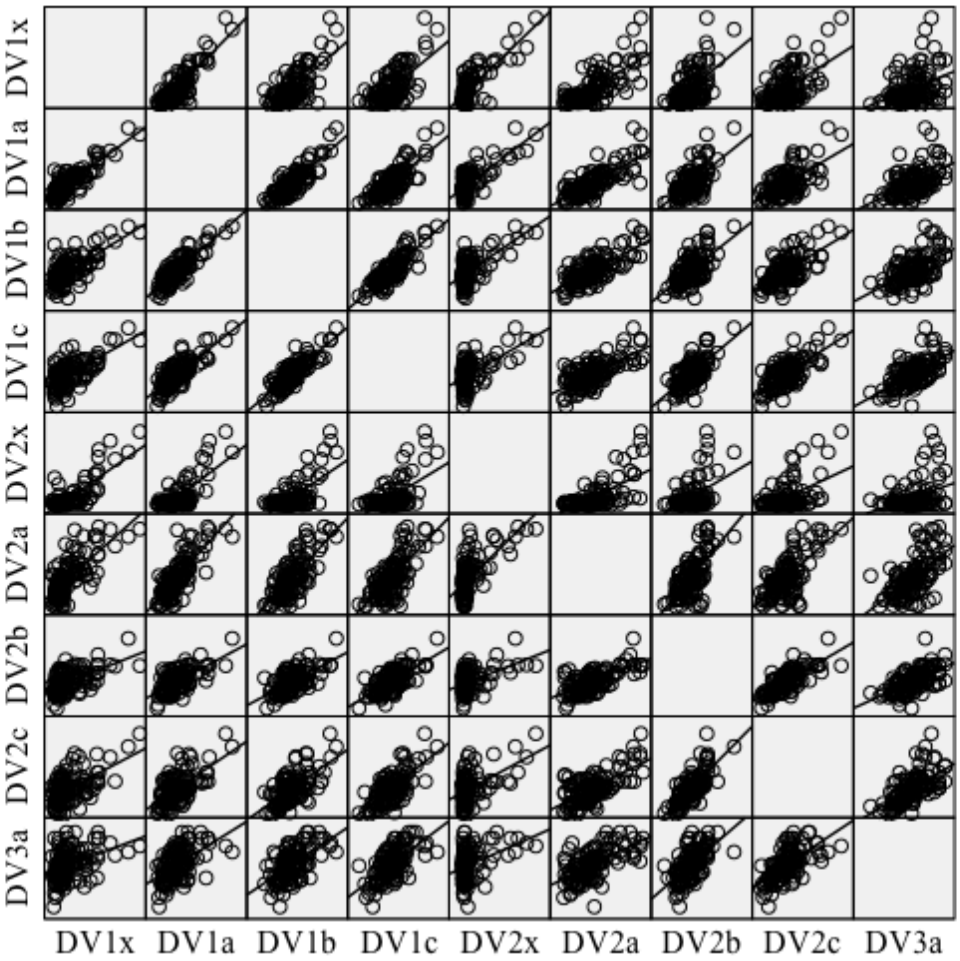
DV1x= Student Scale Score for DIBELS ORF, Beginning of First Grade
DV1a=Student Score for DIBELS ORF, End of First Grade
DV1b=Student Score for DIBELS ORF, End of Second Grade
DV1c=Student Score for DIBELS ORF, End of Third Grade

DV2x= Student Scale Score for STAR, Beginning of First Grade
DV2a=Student Scale Score for STAR, End of First Grade
DV2b=Student Scale Score for STAR, End of Second Grade
DV2c=Student Scale Score for STAR, End of Third Grade

DV3a=Student Scale Score for MCT2, End of Third Grade

APPENDIX C: SCATTERPLOT MATRICES FOR COMPARISON OF LINEARITY

Within-Class Grouping Format

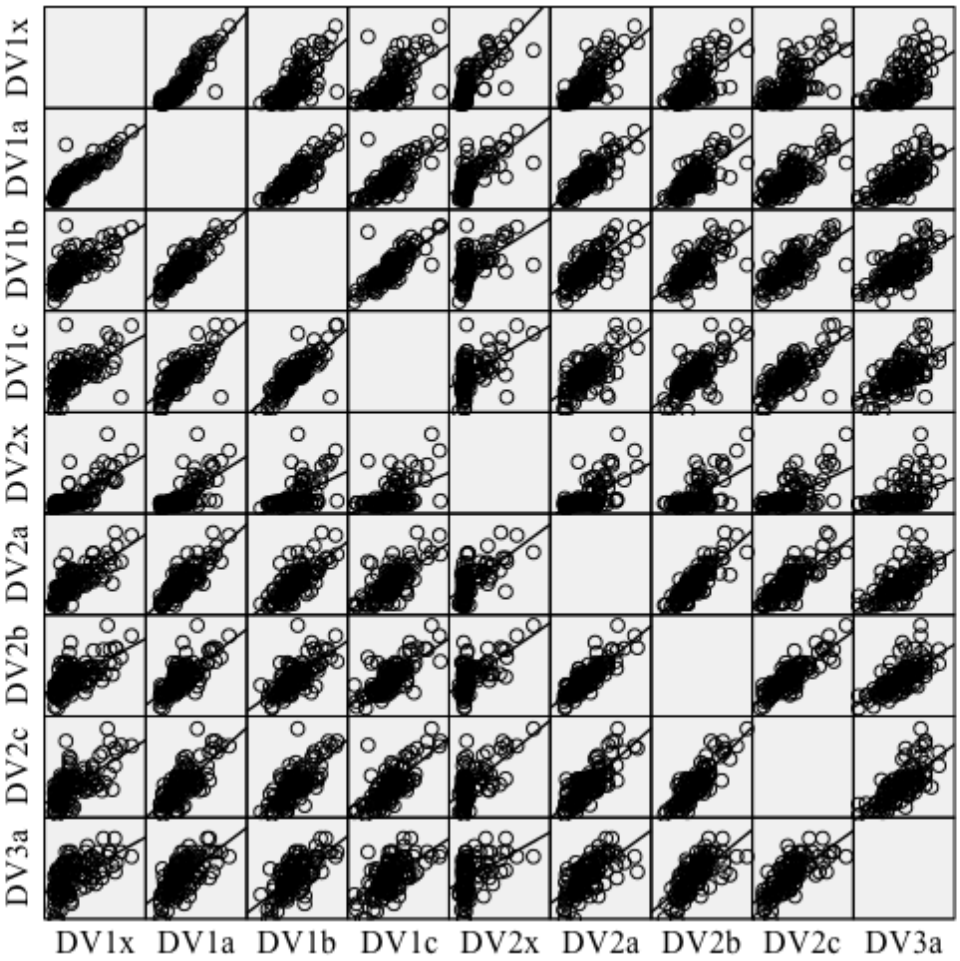


DV1x= Student Scale Score for DIBELS ORF, Beginning of First Grade
DV1a=Student Score for DIBELS ORF, End of First Grade
DV1b=Student Score for DIBELS ORF, End of Second Grade
DV1c=Student Score for DIBELS ORF, End of Third Grade

DV2x= Student Scale Score for STAR, Beginning of First Grade
DV2a=Student Scale Score for STAR, End of First Grade
DV2b=Student Scale Score for STAR, End of Second Grade
DV2c=Student Scale Score for STAR, End of Third Grade

DV3a=Student Scale Score for MCT2, End of Third Grade

Between-Class Grouping Format



DV1x= Student Scale Score for DIBELS ORF, Beginning of First Grade
DV1a=Student Score for DIBELS ORF, End of First Grade
DV1b=Student Score for DIBELS ORF, End of Second Grade
DV1c=Student Score for DIBELS ORF, End of Third Grade

DV2x= Student Scale Score for STAR, Beginning of First Grade
DV2a=Student Scale Score for STAR, End of First Grade
DV2b=Student Scale Score for STAR, End of Second Grade
DV2c=Student Scale Score for STAR, End of Third Grade

DV3a=Student Scale Score for MCT2, End of Third Grade

VITA

S. SUZANNE LIDDELL

EDUCATION

Ed.S., Educational Leadership, Mississippi State University, August 2004

M.Ed., Gifted Studies, Mississippi University for Women, August 1998

B.A., Psychology, Mississippi State University, December 1996

TEACHING AND PROFESSIONAL EXPERIENCE

Director of Federal Programs and Student Registration, June 2011-present
Oxford School District

Principal, June 2006-June 2011
Oxford School District

Assistant Principal, June 2004-June 2006
Oxford School District

Teacher, August 1998-June 2004
Starkville Public Schools
Course: Gifted Education

Graduate Research Assistant, August 1996-July 1998
Mississippi University for Women

HONORS

Administrator of the Year, 2008
Oxford School District
Scholar of the Year, 1996
Department of Psychology
Mississippi State University
Phi Kappa Phi Honor Society, 1995
Golden Key Honor Society, 1994