Using cluster analysis to evaluate the academic performance of demographic homogeneous subsets

Jeffrey N Halsell
University of Nevada, Las Vegas

Follow this and additional works at: https://digitalscholarship.unlv.edu/rtds

Repository Citation
https://digitalscholarship.unlv.edu/rtds/2761

This Dissertation is brought to you for free and open access by Digital Scholarship@UNLV. It has been accepted for inclusion in UNLV Retrospective Theses & Dissertations by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.
USING CLUSTER ANALYSIS TO EVALUATE THE ACADEMIC PERFORMANCE OF DEMOGRAPHIC HOMOGENEOUS SUBSETS

by

Jeffery N. Halsell

Bachelor of Science
University of Nevada, Las Vegas
1994

Master of Arts
University of Nevada, Las Vegas
1999

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Education Degree in Educational Leadership
Department of Educational Leadership
College of Education

Graduate College
University of Nevada, Las Vegas
December 2007
The Dissertation prepared by

Jeffrey N. Halsell

Entitled

Using Cluster Analysis to Evaluate the Criterion

Referenced Test Performance of Demographic Homogenous Subjects

is approved in partial fulfillment of the requirements for the degree of

Doctor of Education in Educational Leadership

Examination Committee Co-Chair

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative
ABSTRACT

Using Cluster Analysis to Evaluate the Academic Performance of Demographic Homogeneous Subsets

by

Jeffrey N. Halsell

Dr. Robert S. McCord, Examination Committee Co-Chair
Professor of Educational Leadership
University of Nevada, Las Vegas
Dr. Chad L. Cross, Examination Committee Co-Chair
Professor of Biostatistics
University of Nevada, Las Vegas

No Child Left Behind (NCLB) requires states to develop, under very specific conditions, a valid and reliable system of measurement that meets certain requirements while neglecting others. One such provision is a system of assessments that holds all schools to the same academic criteria, regardless of circumstance, while at the same time acknowledging an achievement gap exists between certain subpopulations of students. Nevada has developed such a program, which effectively fulfills the requirements of NCLB, and like NCLB, neglects to recognize the unique challenges for those schools housing larger than average populations of lower-performing students.

Historically, when discussing the achievement gap much attention has been placed on social difficulties of certain subpopulations, which may or may not contribute. It is not the intention of this study to examine social issues associated with racial
subgroups or special populations of students. Regardless of the contributing factors, it will suffice to hypothesize a difference in subgroup performance exists. It is simply the intention of this study to examine the performance of schools within homogeneous clusters developed through subgroup membership. It is the belief of this researcher that a school’s dominant student population will have a significant influence on academic performance, which if not considered could result in grave consequences with respect to NCLB. Using cluster analysis, schools were classified based upon dominant student populations and determinations made concerning statistically significant differences in mean reading and mean math CRT scale scores for those schools contained within homogenous clusters.

It was found that although NCLB requires schools to report the academic performance of students belonging to subgroup’s American Indian, Asian, Hispanic, African American, White, IEP, LEP, and FRL, the only subgroups that provide valid and measureable results were Asian, African American, Hispanic, White, and IEP. Further, schools did demonstrate significant differences in mean reading and mean math scale scores with select schools performing significantly above expectations, certain schools performing significantly below expectations, and many demonstrating no significant difference relative to similar populations located in homogeneous clusters.
TABLE OF CONTENTS

ABSTRACT ......................................................................................................................................... iii
LIST OF TABLES ............................................................................................................................ viii
LIST OF FIGURES ............................................................................................................................ ix
ACKNOWLEDGEMENTS .............................................................................................................. xi

CHAPTER 1 OVERVIEW ............................................................................................................. 1
   Introduction ............................................................................................................................... 1
   The Nevada Plan ......................................................................................................................... 3
   Statement of the Problem ......................................................................................................... 6
   Purpose of the Study ............................................................................................................... 7
   Research Questions ................................................................................................................... 7
   Conceptual Framework .......................................................................................................... 11
   Summary of Methodology ...................................................................................................... 12
   Sources of Data ...................................................................................................................... 13
   Definition of Terms ................................................................................................................. 13
   Summary ................................................................................................................................ 15

CHAPTER 2 REVIEW OF THE LITERATURE ........................................................................... 16
   Introduction ............................................................................................................................... 16
   Adequate Yearly Progress ....................................................................................................... 17
   Measurement Validity and Reliability .................................................................................... 20
   Large Scale Testing Programs .............................................................................................. 22
   Sampling Error ....................................................................................................................... 23
   Subgroup Size .......................................................................................................................... 28
   Subgroup Performance and Diverse Schools ......................................................................... 30
   Students with Limited English Proficiency ........................................................................ 34
   Summary ................................................................................................................................ 38

CHAPTER 3 RESEARCH DESIGN AND METHODOLOGY ............................................... 41
   Introduction ............................................................................................................................... 41
   Statement of the Problem ....................................................................................................... 42
   Purpose of the Study .............................................................................................................. 43
   Research Questions ................................................................................................................ 43
   Research Design ...................................................................................................................... 47
   Data and Data Sources ............................................................................................................ 52
LIST OF TABLES

Table 1  Subgroup N-counts and Percentages .................................................................56
Table 2  Initial Cluster Distributions ..............................................................................59
Table 3  Initial Cluster Centroids ..................................................................................60
Table 4  Pearson’s Correlation Coefficient of Subgroup Z-score ...............................78
Table 5  Asian, LEP, and FRL Cluster Distributions .......................................................85
Table 6  Asian, LEP, and FRL Cluster Centroids ...........................................................86
Table 7  Hispanic, LEP, and FRL Cluster Distributions .................................................92
Table 8  Hispanic, LEP, and FRL Cluster Centroids ......................................................93
Table 9  African American, LEP, and FRL Cluster Distributions ...............................99
Table 10 African American, LEP, and FRL Cluster Centroids ................................100
Table 11 White, LEP, and FRL Cluster Distributions ....................................................105
Table 12 White, LEP, and FRL Cluster Centroids .........................................................106
Table 13 Phase II Cluster Distributions ........................................................................108
Table 14 Phase II Cluster Centroids .............................................................................108
Table 15 African American ANOVA Summary Table ................................................123
Table 16 Levene’s Test of Homogeneity of Variance for African American.............124
Table 17 Welch Robust Tests of Equality of Means for African American .............124
Table 18 Asian ANOVA Summary Table ...................................................................129
Table 19 Levene’s Test of Homogeneity of Variance for Asian .............................129
Table 20 Hispanic ANOVA Summary Table ...............................................................134
Table 21 Levene’s Test of Homogeneity of Variance for Hispanic ......................134
Table 22 Welch Robust Tests of Equality of Means for Hispanic .........................134
Table 23 White ANOVA Summary Table ................................................................143
Table 24 Levene’s Test of Homogeneity of Variance for White ..........................143
Table 25 Welch Robust Tests of Equality of Means for White .............................143
Table 26 IEP ANOVA Summary Table .....................................................................149
Table 27 Levene’s Test of Homogeneity of Variance for IEP ..............................150
Table 28 Welch Robust Tests of Equality of Means for IEP .................................150
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>American Indian within Cluster Variation</td>
<td>61</td>
</tr>
<tr>
<td>2</td>
<td>American Indian Clusterwise Importance</td>
<td>62</td>
</tr>
<tr>
<td>3</td>
<td>Asian within Cluster Variation</td>
<td>63</td>
</tr>
<tr>
<td>4</td>
<td>Asian Clusterwise Importance</td>
<td>64</td>
</tr>
<tr>
<td>5</td>
<td>Hispanic within Cluster Variation</td>
<td>65</td>
</tr>
<tr>
<td>6</td>
<td>Hispanic Clusterwise Importance</td>
<td>66</td>
</tr>
<tr>
<td>7</td>
<td>African American within Cluster Variation</td>
<td>67</td>
</tr>
<tr>
<td>8</td>
<td>African American Clusterwise Importance</td>
<td>68</td>
</tr>
<tr>
<td>9</td>
<td>White within Cluster Variation</td>
<td>69</td>
</tr>
<tr>
<td>10</td>
<td>White Clusterwise Importance</td>
<td>70</td>
</tr>
<tr>
<td>11</td>
<td>IEP within Cluster Variation</td>
<td>71</td>
</tr>
<tr>
<td>12</td>
<td>IEP Clusterwise Importance</td>
<td>72</td>
</tr>
<tr>
<td>13</td>
<td>LEP within Cluster Variation</td>
<td>73</td>
</tr>
<tr>
<td>14</td>
<td>LEP Clusterwise Importance</td>
<td>74</td>
</tr>
<tr>
<td>15</td>
<td>FRL within Cluster Variation</td>
<td>75</td>
</tr>
<tr>
<td>16</td>
<td>FRL Clusterwise Importance</td>
<td>76</td>
</tr>
<tr>
<td>17</td>
<td>Asian, with LEP and FRL within Cluster Variation</td>
<td>80</td>
</tr>
<tr>
<td>18</td>
<td>Asian, with LEP and FRL Clusterwise Importance</td>
<td>81</td>
</tr>
<tr>
<td>19</td>
<td>LEP, with Asian and FRL within Cluster Variation</td>
<td>82</td>
</tr>
<tr>
<td>20</td>
<td>LEP, with Asian and FRL Clusterwise Importance</td>
<td>83</td>
</tr>
<tr>
<td>21</td>
<td>FRL, with Asian and LEP within Cluster Variation</td>
<td>84</td>
</tr>
<tr>
<td>22</td>
<td>FRL, with Asian and LEP Clusterwise Importance</td>
<td>85</td>
</tr>
<tr>
<td>23</td>
<td>Hispanic, with LEP and FRL within Cluster Variation</td>
<td>87</td>
</tr>
<tr>
<td>24</td>
<td>Hispanic, with LEP and FRL Clusterwise Importance</td>
<td>88</td>
</tr>
<tr>
<td>25</td>
<td>LEP, with Hispanic and FRL within Cluster Variation</td>
<td>89</td>
</tr>
<tr>
<td>26</td>
<td>LEP, with Hispanic and FRL Clusterwise Importance</td>
<td>90</td>
</tr>
<tr>
<td>27</td>
<td>FRL, with Hispanic and LEP within Cluster Variation</td>
<td>91</td>
</tr>
<tr>
<td>28</td>
<td>FRL, with Hispanic and LEP Clusterwise Importance</td>
<td>92</td>
</tr>
<tr>
<td>29</td>
<td>African American, with LEP and FRL within Cluster Variation</td>
<td>94</td>
</tr>
<tr>
<td>30</td>
<td>African American, with LEP and FRL Clusterwise Importance</td>
<td>95</td>
</tr>
<tr>
<td>31</td>
<td>LEP, with African American and FRL within Cluster Variation</td>
<td>96</td>
</tr>
<tr>
<td>32</td>
<td>LEP, with African American and FRL Clusterwise Importance</td>
<td>97</td>
</tr>
<tr>
<td>33</td>
<td>FRL, with African American and LEP within Cluster Variation</td>
<td>98</td>
</tr>
<tr>
<td>34</td>
<td>FRL, with African American and LEP Clusterwise Importance</td>
<td>99</td>
</tr>
<tr>
<td>35</td>
<td>White, with LEP and FRL within Cluster Variation</td>
<td>100</td>
</tr>
<tr>
<td>36</td>
<td>White, with LEP and FRL Clusterwise Importance</td>
<td>101</td>
</tr>
<tr>
<td>37</td>
<td>LEP, with White and FRL within Cluster Variation</td>
<td>102</td>
</tr>
<tr>
<td>38</td>
<td>LEP, with White and FRL Clusterwise Importance</td>
<td>103</td>
</tr>
<tr>
<td>39</td>
<td>FRL, with White and LEP within Cluster Variation</td>
<td>104</td>
</tr>
<tr>
<td>40</td>
<td>FRL, with White and LEP Clusterwise Importance</td>
<td>105</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

This degree was made possible through the support of many, a true collaborative effort. Completing such a task requires dedication and perseverance, which further requires a varied collection of support, understanding, and confidence from those closest to the project.

I would like to begin by recognizing the support and encouragement of my co-chairmen, Dr. Robert S. McCord and Dr. Chad Cross. Dr. McCord, who also serves as co-director of the UNLV Center for Educational Policy Studies, provided invaluable expertise and advice on educational policy while Dr. Cross’s knowledge of statistics, and especially his expertise with applying novel techniques to non-standard data sets, enabled me to effectively venture into untested reporting procedures. Without their assistance, this project would have never been completed. I would also like to recognize committee member Dr. James Crawford, whose untiring editing and effective questioning techniques kept this project on track, and Dr. Lawrence R. Walker, who unselfishly donated his valuable time and effort to serve as my graduate representative. In addition, sincere thanks to Dr. Rodney Young, whose technical advice and encouragement assisted in keeping this project on task. My sincere gratitude goes out to this group, whose patience and encouragement brought this project to completion.

I would also like to acknowledge the support and encouragement of my children Jessecca, Adam, Jeremiah, and Carey. Their never-ending requests for progress reports made it necessary to move forward and report advancement. Besides, it is difficult to
expect your children to complete tasks if you yourself cannot. Also important was the pride my sister Teresa Thompson felt toward my goal of completing a doctorate. Her faith in me was invaluable. In addition, my aunt Pauline Proffitt has supported and encouraged my education from the time I was a child. Without her, I doubt I would have ever achieved such a goal. And my father-and mother-in-law, Bob and Evelyn Tallent, have always been there to provide whatever type of support was necessary. They could never truly understand the value of their contributions.

Finally, I would like to recognize the contributions of my wife Nita. From the beginning, she has been there to provide support and encouragement. When things were good, she was there to celebrate with me, and when things were bad, she was behind me, pushing me to get back to work and overcome the obstacle. While many have been there to encourage and help me, she has been the one that was there at the beginning and supported me throughout. Words could never express the gratitude I feel toward her.
CHAPTER 1

OVERVIEW

Introduction

The Nevada Criterion Referenced Testing (CRT) program, mandated by legislation in 1999 and first piloted in Math and Reading during the 2000-2001 school year, was designed to measure student academic achievement in Nevada State Content and Performance Standards (U.S. Department of Education, 2005). Results from the CRT testing program were used to monitor individual as well as school and district performance on a pre-determined criterion of correct responses. Grades three through eight CRT achievements were evaluated using four levels of ability; emergent/developing, approaches standards, meets standards, and exceeds standards. A 100-500 scale score system was used with scores greater than 300 representing meets or exceeds standards. Students scoring 300 or greater on the Nevada CRT examination were designated as proficient while students scoring less than 300 were classified as emergent/developing or approaches standards and designated non-proficient.

In January, 2002, The Elementary and Secondary Education Act was reauthorized through passage of The No Child Left Behind Act (NCLB) Act (Public Law 107-110). The state of Nevada, in response to federal requirements associated with NCLB, realigned its state accountability guidelines through passage of Senate Bill 1 during the
19th special session of the Nevada Legislature (SB1, 2003). To fulfill the requirements associated with NCLB and SB1, the state of Nevada, each school district, and all public school sites were evaluated annually on how well their student populations advanced toward a predetermined level of academic achievement. Adequate Yearly Progress (AYP), a term used by the United States Department of Education (USDOE) to describe annual growth, was a prescriptive measure designed to reach 100% proficiency for all students on or before the 2013-2014 school year. Assessments were administered in Reading and Mathematics for grades three through eight using the Nevada Department of Education (NDE) Criterion Referenced tests, grades five and eight using the NDE Writing assessment, grades ten through twelve using the Nevada High School Proficiency Exam (NHSPE) in Reading and Mathematics, and grade eleven through twelve with the NHSPE Writing examination.

Section 1001.3 of the No Child Left Behind Act of 2001 (PL 107-110) recognized an achievement gap existed between subclasses of students and specifically stated the purpose NCLB, among other things, was to:

"Close the achievement gap between high- and low-performing children, especially the achievement gaps between minority and nonminority students, and between disadvantaged children and their more advantaged peers."

Identifying and reducing or eliminating an achievement gap between subclasses of students should be the goal of all educators. Unfortunately, the Consolidated State Application Accountability Workbook, published by the U.S. Department of Education (U.S. Department of Education, 2005) to provide guidance when designing a state accountability system to meet the requirements of NCLB, called for an accountability...
system that holds all schools to the same criteria, regardless of circumstance. No provisions were included for addressing the achievement gap that already existed, which resulted in schools enrolling majority populations of so called high-performing children enjoying the rewards that accompany high-achieving schools while those housing a majority of low-performing disadvantaged students suffered the sanctions associated with not making Adequate Yearly Progress. It is the purpose of this study to examine an alternative method for evaluating Adequate Yearly Progress that considers achievement relative to population served.

The Nevada Plan

NCLB requires all grade appropriate students to participate in state-mandated testing with a minimum of 95% participation required to meet standards. In addition to evaluating and reporting the percentage of students participating, NCLB further requires that schools report the rate of proficiency for those students enrolled in a particular school or district for the full academic school year. The full academic school year had been defined as continuous enrollment from the official count day, which occurs on or about the third week of September, through the specified testing window, which for CRT testing occurred in late March/early April. Therefore, when reporting the performance of a particular school, NCLB requires and reports the participation rate for all grade appropriate students enrolled during the testing period, regardless of date of enrollment, while proficiency rate would include only those students enrolled for the full academic year. A similar measure is used to determine if a student’s academic result should be
aggregated into district level reports, with students enrolled in the district after count day exempt from having their academic performance aggregated into district level reports.

In addition to evaluating the participation and proficient status for all eligible students at the district and school level, results are further disaggregated into student subgroups. The participation rate for subgroups with greater than 20 members is calculated using the proportion of students participating with at least 95% needed to meet standards. For subpopulations with less than 20 members, NCLB required all participate less one, or \( n - 1 \). Proficiency rates were calculated for subgroups with 25 or more members using the proportion of students that had been enrolled for the full academic year and were meeting or exceeding standards. And finally, the academic performances for subgroup populations with fewer than 25 members were not to be reported, regardless of the performance of its members.

Reported subgroups include all major ethnic groups (American Indian/Alaskan Native, Asian/Pacific Islander, Hispanic, African American, and White), students with Individual Education Plans (IEP), students with Limited English Proficiency (LEP), and students eligible for Free or Reduced Lunches (FRL). All eligible students will be in the school or district participation rate with those enrolled for the full academic year in the school or district academic performance report. One exception is a small percentage of students not reported in an ethnic group, most will also be members of, and therefore evaluated in, a unique ethnic group. This reporting of ethnic groups results in the majority of student participation rates and academic results reported twice. And, besides being evaluated in the school or district report and ethnic group, many students are also members of, and therefore evaluated in, one or more special programs. This method of
reporting by subgroup population results in an academically eligible student’s performance evaluated at least twice and up to as many as five times.

The United States Department of Education requests that all states receiving funds through the Title I program participate in NCLB. Under NCLB, annual reports must summarize student performance across 37 separate areas, each area evaluated independently, using a simple pass-fail criterion. Each school must have at least 95% participation in English/Language Arts (ELA) and math testing, with subgroups enrolling less than 20 students allowed a maximum of one non-participant. Participation rate is calculated for the school as whole, students that are enrolled as American Indian/Alaskan Native, Asian/Pacific Islander, Hispanic, African American, or White, students with an IEP, students that are LEP, and students that qualify for FRL. Schools must also reach a pre-determined percentage of proficiency in ELA and math, using only those students that have been enrolled for the full academic year. Evaluations are made for the school as a whole, each of the five previously mentioned ethnic groups with 25 or more students, and each of the three special populations with 25 or more members. And lastly, schools must adequately perform one other indicator (OI) at the school level. For elementary and middle schools, the OI measure is average daily attendance of 90% or more with high schools requiring a graduation rate of 50% or more.

While participation rate, graduation rate, and average daily attendance are important variables to consider, the primary intention behind NCLB legislation is to narrow and, if possible, eliminate the achievement gap that exists between discrete ethnic and special populations (Kim & Sunderman, 2005). Section 1001 of the No Child Left Behind Act of 2001 (PL 107-110) specifically states:
"The purpose of this title is to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging State academic achievement standards and state academic assessments. This purpose can be accomplished by —

1. ensuring that high-quality academic assessments, accountability systems, teacher preparation and training, curriculum, and instructional materials are aligned with challenging State academic standards so that students, teachers, parents, and administrators can measure progress against common expectations for student academic achievement;

2. meeting the educational needs of low-achieving children in our Nation's highest-poverty schools, limited English proficient children, migratory children, children with disabilities, Indian children, neglected or delinquent children, and young children in need of reading assistance;

3. closing the achievement gap between high- and low-performing children, especially the achievement gaps between minority and nonminority students, and between disadvantaged children and their more advantaged peers;”

Statement of the Problem

NCLB mandates all schools be evaluated annually using the same criteria without consideration for unique factors, such as larger than average low-performing or special needs populations. While designed to eliminate the achievement gap between high and low performing students, NCLB, in its current design, singles out for sanctions and excludes from rewards schools serving racially diverse, special needs student populations
(Kane, Staiger & Geppert, 2001; Kim & Sunderman, 2005). Considering the high stakes associated with failing to meet Adequate Yearly Progress, is the statute mandating schools are evaluated using a standardized approach, without first considering the school’s unique population and achievement relative to others with a similar demographic profile, sound federal policy?

Purpose of the Study

It is the purpose of this study to evaluate the efficacy of NCLB using Nevada standards based CRT results to determine if overall academic performance and annual progress can be evaluated with respect to unique demographic characteristics. It is hypothesized that a school’s dominant student population will have a significant influence on academic performance and would enhance results if considered. By using cluster analysis, schools were classified using dominant student populations and decisions will be made about achievement using statistically significant differences in mean reading and mean math CRT scale scores. This study should determine if academic performance is influenced by demographic profile, and, if so, which schools sharing similar demographic profiles perform above or below expectations.

Research Questions

The research questions for this study are:

1. Which demographic variables, as recognized through NCLB, generate unique and homogeneous clusters of five or more schools?
a. Demographic variables generate unique and homogeneous clusters when examining the American Indian/Alaskan Native subpopulation.
b. Demographic variables generate unique and homogeneous clusters when examining the Asian subpopulation.
c. Demographic variables generate unique and homogeneous clusters when examining the Hispanic subpopulation.
d. Demographic variables generate unique and homogeneous clusters when examining the African American subpopulation.
e. Demographic variables generate unique and homogeneous clusters when examining the White subpopulation.
f. Demographic variables generate unique and homogeneous clusters when examining the IEP subpopulation.
g. Demographic variables generate unique and homogeneous clusters when examining the LEP subpopulation.
h. Demographic variables generate unique and homogeneous clusters when examining the FRL subpopulation.

2. Does a statistically significant difference in mean reading CRT scale score exist within homogeneous clusters of five or more schools?
   a. If the subgroup American Indian/Alaskan Native generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.
b. If the subgroup Asian generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

c. If the subgroup Hispanic generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

d. If the subgroup African American generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

e. If the subgroup White generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

f. If the subgroup IEP generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

g. If the subgroup LEP generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

h. If the subgroup FRL generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

3. Does a statistically significant difference in mean math CRT scale score exist within homogeneous clusters of five or more schools?
a. If the subgroup American Indian/Alaskan Native generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

b. If the subgroup Asian generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

c. If the subgroup Hispanic generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

d. If the subgroup African American generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

e. If the subgroup White generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

f. If the subgroup IEP generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

g. If the subgroup LEP generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.
h. If the subgroup FRL generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

**Conceptual Framework**

The conceptual framework for this study involves considering alternative methods when evaluating the academic performance of state and local educational agencies. The current practice of determining AYP, using standardized methodologies, does not take into consideration unique needs and characteristics of districts and schools. The Nevada NCLB workbook references identical methods for measuring the academic achievement of students in Clark County, Nevada, the fifth largest urban school district in the United States enrolling approximately 310,000 students, as it does Esmeralda County, Nevada, a rural district in west-central Nevada enrolling approximately 68 students. Similar inner-district imbalances occur when the academic performance of upscale, affluent schools is compared to schools located in inner-city, high-risk neighborhoods.

The primary goal of NCLB is to narrow, and if possible, eliminate the achievement gap between minority and non-minority students. Unfortunately, current federal policy does not take into consideration, and thereby compensate for, pre-existing achievement gaps. It is the purpose of this study to investigate an alternative method that will recognize academic success or failure with respect to pre-existing conditions and unique demographic profiles.
Summary of Methodology

The proportion of ethnic and special population students enrolled at individual school sites were standardized, using a mean of zero and standard deviation of one (μ=0, σ=1), with cluster membership assigned using the TwoStep cluster algorithm. The TwoStep method, designed to handle continuous and categorical variables, determined cluster membership through minimizing the distance between data points and corresponding cluster centroids using log-likelihood distances. The initial TwoStep cluster procedure generated eight unique clusters, corresponding to the eight standardized variables (American Indian, Asian, Hispanic, African American, White, IEP, LEP, FRL), with the number of clusters equal to the number of variables. If the eight dependent variables did not generate eight unique homogeneous subsets, or when dependent variables demonstrated strong multicollinearity, the dependent variables were reduced to eliminate ineffective or irrelevant variables and the TwoStep cluster procedure repeated with n-less clusters.

When \( n \) homogeneous clusters were generated, each statistically significant to a single variable, a One-Way Analysis of Variance (ANOVA) procedure was performed on each cluster to look for differences in mean reading and mean math CRT scale scores. The ANOVA tested the null hypothesis that there were no statistically significant differences in mean reading or mean math CRT scale scores for schools located within each unique cluster. When a statistically significant difference in the mean reading or mean math CRT scale score was found, and the cluster schools demonstrated equal variances, Tukey’s post-hoc test was used to identify which school’s performance on the
reading or math CRT was significantly different from the others. In such cases where variances were unequal, Tamhane's T2 post hoc procedure was used.

Sources of Data

The data source used for this study was taken from the Clark County School District grades three through five reading and math CRT assessment results collected during the spring, 2006 administration. The data file included student level school location name and code, ethnic and special population membership, enrollment information to identify enrollment before or after count day, reading and math CRT raw scores, reading and math CRT scale scores, and level of academic performance. This data set fulfilled all requirements for test reliability and validity.

Definition of Terms

Analysis of Variance (ANOVA) – Statistical tool used to identify the relationship between a response variable and one or more explanatory variables (Neter, et. al., 1996).

Average Daily Attendance (ADA) - Refers to the average percentage of students present in a school over the course of the year.

Adequate Yearly Progress (AYP) - An accountability system prescribed by the federal government to determine if schools are making process toward narrowing the achievement gap and ensuring all students are proficient in the areas of mathematics and English Language Arts by the 2013-2014 school year.

Cluster Analysis - A multivariate statistical procedure that reorganizes a data set into relatively homogeneous groups (Aldenderfer & Blashfield, 1984).
Construct Validity - The extent to which variables accurately measure the constructs of interest (Vogt, 1999).

Criterion-Referenced Test (CRT) - Refers to reading, mathematics, and science tests in Nevada based on state standards.

English Language Arts (ELA) - ELA assessments include reading and writing.

Free or Reduced Price Lunch (FRL) - Refers to students qualifying for free or reduced price lunches. Commonly used as a proxy for socio-economic status.

Individualized Education Plan (IEP) - Refers to students who receive special educational services due to a learning disability or cognitive deficit.

Limited English Proficient (LEP) - Refers to students who are learning English as a second language and qualify for English language learner (ELL) services.

No Child Left Behind (NCLB) - The commonly used name to refer to house referendum 1, the 2001 reauthorization of the Elementary and Secondary Education Act.

Nevada Department of Education (NDE) - The state of Nevada educational agency.

Other Indicator (OI) - Refers to additional criteria used to evaluate schools. In Nevada other indicators include average daily attendance for elementary and middle schools and graduation rate for high schools.

Percent Above Cut (PAC) - Refers to the percentage of students scoring at or above proficient level on state standardized tests.

Socio Economic Status (SES) - Reflects the economic standing of students' parents or primary providers. Commonly derived from students' eligibility for free or reduced price lunches.
Summary

As mandated through NCLB and SB1 the state of Nevada, each school district in the state of Nevada, and all public school sites located within the state of Nevada will be evaluated annually on how well their student populations advance toward a predetermined level of academic achievement with the eventual goal of 100% proficiency on or before the 2013-2014 school year. While designed to narrow, and if possible, eliminate the achievement gap, states and districts are finding that certain elements of the law make it difficult, if not impossible, to meet the requirements. Fluctuations resulting from cohort abilities, as well as variations in ability level between homogeneous and racially diverse schools, single out and punish those schools serving racially diverse and large special needs populations. It is the purpose of this study to examine the efficacy of considering unique demographic characteristics when evaluating overall academic performance and annual progress.
CHAPTER 2

REVIEW OF THE LITERATURE

Introduction

The No Child Left Behind Act of 2002, the most recent reauthorization of the Elementary and Secondary Act of 1965 (NCLB; Public Law No. 107-110, 115 Stat. 1425, 2002), was designed to guarantee all students receive a fair, equitable, high-quality education. One major provision of NCLB is an annual measure of Adequate Yearly Progress (AYP) in English/language arts and mathematics. State, district and school AYP measurements are reported for the group as a whole as well as for major ethnic groups, which include American Indian /Alaskan Native, Asian/Pacific Islander, Hispanic, African American, and White, students with Individual Education Plans (IEP), students with Limited English Proficiency (LEP), and students eligible for Free or Reduced Lunches (FRL). In addition to reports of annual rate of proficiency, states, districts and schools also report percentage of students participating, average daily attendance and graduation rate, where applicable.

To meet requirements outlined through NCLB, the state of Nevada utilized an in-place system of assessment. The Nevada Criterion Referenced Assessment (CRT) program, which was initially mandated in 1999 and piloted for math and reading during the 2000-2001 school year (U.S. Department of Education, 2005), was adopted in 2002
as the measurement system to determine elementary and middle school AYP. The Nevada CRT program evaluated how well an individual student, school or district performed using a pre-determined criterion of correct responses. Grades three through eight CRTs were evaluated using a 100-500 scale score system with scale scores of 300 or greater representing meets or exceeds standards.

Adequate Yearly Progress

As mandated through NCLB, Nevada’s definition of AYP requires all students to be proficient in English/Language Arts and mathematics by the 2013-2014 academic year (Public Law 107-110). To determine the baseline rate of proficiency, Nevada used the school percentile method, which involved ranking schools in terms of subject area proficiency while cumulating enrollment up to and including the 20th percentile (Marion, et. al., 2002). The rate of proficiency for the school at the statewide 20th cumulative percentile was the baseline performance rate for Nevada in that particular subject area with a separate measure established for English/Language Arts and Math. Future targets were established, using a tiered method, by subtracting from 100% the subject area baseline and dividing that value by six, establishing six equal interval increases to an eventual 100% proficiency at the 2013-2014 school year (LaMarca, 2005). While this technique guaranteed 80% of all school-wide measurements would achieve Adequate Yearly Progress that first measurement year, the method did not consider subgroup performance, which is also required for a school to make Adequate Yearly Progress.

Schools can fail to achieve AYP in three separate ways. The school could fail to achieve participation and/or proficiency requirements in any one of the nine separate
areas for English/Language Arts, the school could also fail to achieve participation and/or
proficiency requirements in any one of the nine separate areas for math, or the school
could fail to achieve the school-wide other indicator measurement of 90% ADA for
elementary and middle school or 50% graduation rate for high school. The nine separate
areas for failing AYP in ELA and Math are the school-wide measure, any one of five
reported ethnic groups, or any one of three special populations that includes IEP, LEP, or
FRL (LaMarca, 2005).

The penalty for not making Adequate Yearly Progress has been defined by
failing to make AYP for two continuous years in one or more areas of ELA, math, or
school-wide other indicator, the school is identified as “in need of improvement year 1”
(N1) and will:

- Use federal funding to acquire technical assistance to improve
  achievement
- Develop a school improvement plan
- Offer school choice to all students
- Provide transportation for those that choose to attend a different school

After three years of failing to achieve adequate yearly progress in the same area the
school will be designated N2 and will:

- Continue to use federal funding to acquire technical assistance to improve
  achievement
- Develop a new school improvement plan
- Continue to offer school choice to all students
• Continue to provide transportation for those that choose to attend a
different school
• Use Title I funding to purchase and provide transportation to student
  supplemental educational services

If the school fails to achieve AYP in the same subject area for four continuous years, the
school will be designated N3 and will:

• Continue to use federal funding to acquire technical assistance to improve
  achievement
• Develop a new school improvement plan
• Continue to offer school choice to all students
• Continue to provide transportation for those that choose to attend a
different school
• Continue to use Title I funding to purchase, and provide transportation to,
  student supplemental educational services
• Implement district corrective actions that may include a change in school
  leadership, staff or programming

If a school fails AYP in the same area for five continuous years (N4), the school should
prepare for complete restructuring.

One final element of NCLB, the safe harbor stipulation, was included to insure
any school or subgroups that did not achieve the targeted rate of proficiency but could
demonstrate significant positive growth were not penalized. This provision allowed any
subgroup to pass, regardless of rate of proficiency, if that subgroup could demonstrate a
10% or more reduction in non-proficient students and satisfy the subgroup other indicator
provision. For example, if the IEP subgroup could decrease the percentage of non-proficient students by at least 10%, and the average daily attendance for the IEP subgroup was adequate, the IEP subgroup would make adequate yearly progress regardless of the original rate of proficiency (Coladarci, 2003; Keegan, Orr & Jones; 2002).

Kane, Staiger & Geppert (2001) introduced two important points that deserve consideration when evaluating performance. First, by requiring periodic increases to student proficiency, NCLB overlooks natural fluctuations in year-to-year student performance. Cohorts of students can be very sensitive to the talents or rowdiness of a particular group and may increase or decrease rate of proficiency based upon group dynamics, not instructional practices. Therefore, reliance on the safe-harbor provision is in effect a reliance on cohort stability. And secondly, the bill fails to recognize its impact on racially diverse schools. Schools with homogeneous populations report the performance of a single racial subgroup, whereas schools with diverse populations report the performance of many racial subgroups. Viewing each reported subgroup as an independent evaluation increases a school’s probability of not meeting standards. Taking into consideration the high correlation between race and special populations, minority subgroups are more likely to belong to economically depressed and limited English proficient subgroups, thereby increasing their propensity to not meet requirements through NCLB (Kane, Staiger & Gippert, 2001; Orfield, 1996; Orfield & Lee; 2005).

**Measurement Validity and Reliability**

In many states, including Nevada, the standards-setting process was carried out without knowledge of future sanctions tied to NCLB. Because each state was given the
flexibility to develop their own standards-based assessments, levels of difficulty tend to vary from state to state. In 2001, Mississippi reported 39% of their grade eight students were performing at proficient or greater in state mandated criterion referenced testing, Louisiana reported a mere 7% proficient, and Texas an extraordinary 92% proficient (Linn, Baker & Betebenner, 2002). As this example illustrates, the definition of proficient can vary with states having low expectations more likely to achieve Adequate Yearly Progress than those with ambitious testing criterion and high performance standards (Linn, Baker & Betebenner, 2002).

Another area of concern is volatility resulting from measurement error. Measurement error is generated from two primary sources; sampling error caused by testing a different group of students each year, and measurement error resulting from environmental influences such as a dog barking outside, feeling ill, etc. (Kane, Staiger & Geppert, 2001; Hill & DePascale, 2003; Linn & Haug, 2002). Cross-sectional measurement, or the comparison of current student populations to groups evaluated in previous years, is a major source of sampling error. Cohort instability, such as having a low performing group of students replaced by an above-average group, can produce fluctuations in year-to-year student performance. The influence of sampling error on AYP measurements could produce different classifications each year with no significant changes to instructional practices (Kane, Staiger & Geppert, 2001; Hill & DePascale, 2003).

Also significant to test volatility is measurement error associated with subgroup sample size. Small sample groups have the potential to produce large variations simply due to number of measurements while larger subgroups tend to provide a more stable
result. For example, a subgroup with 10 members could realistically achieve 100% proficiency as well as 0% proficiency whereas a subgroup of 1000 would be unlikely to report all of its members proficient or non-proficient. Sampling error is a function of the square root of samples, meaning every time the number of students increases by a factor of four the sampling error is halved (Hill & DePascale, 2003). For accountability purposes, Nevada has determined that subgroups with less than 25 members are too small to provide statistically reliable results and will therefore not report their performance, regardless of outcome.

**Large Scale Testing Programs**

Central to NCLB is the annual measure and public report of the academic progress of all students. The Standards for Educational and Psychological Testing (2004) define a test as “a set of tasks designed to elicit or a scale to describe examinee behavior in a specified domain, or a system for collecting samples of an individual’s work in a particular area (pg. 25).” According to Haertel (1999), large scale assessments programs serve four major functions:

1. Provide analysis to evaluate accountability programs and compare/evaluate schools and districts.
2. Publicly highlight educational concerns and issues.
3. Influence educational practice, curriculum and instruction.
4. Stimulate effort on the part of school administrators, teachers, and students.

Linn (2000) views large scale assessment from another perspective. In his opinion, the appeal for selecting large scale assessment programs as an agent for change is because:
1. Compared to changes in instructional practices, assessment programs are relatively inexpensive.

2. Assessment programs can be externally mandated, which is much easier than exacting change in the classroom.

3. The implementation of a large scale assessment program is relatively quick.

4. Results are visible and can be publicly reported.

Linn further points out most new large-scale assessment programs experience increases in the first few years, with or without any real academic gains. These short term gains, largely due to construct related error through students and staff simply becoming familiar with the assessment mechanics, can provide quick and positive results.

**Sampling Error**

As indicated earlier, sampling error will cause student results to vary from year to year, even if the curriculum, instruction, and community the students come from remain constant. Hanushek & Raymond (2002) view the simplest model of student achievement as:

\[
\text{Achievement} = \text{school} + \text{others}
\]

Where:

\[
\text{Others} = \text{ability} + \text{family} + \text{peers} + \text{history} + \text{error}
\]

It is their belief a variety of factors contribute to a school's overall success or failure. They conclude some aggregation of the assessment can be used to evaluate the effectiveness of the school, but much will depend on factors that lie outside the school environment.
The assessment error component, which could be positive or negative, will include sampling error, measurement error, equating error, and systematic error (Arce-Ferrer, Frisbie, and Kolen, 2002). Yen (1997) defines the additive components of the PAC standard error for school $i$ as:

$$
SE^2(PAC)_i = \frac{\sigma_f^2}{F_i} + \frac{\sigma_{sf}^2}{F_i} + \frac{\sigma_p^2}{n_{ri}} \cdot \frac{N_{ri} - n_{ri}}{N_{ri}} + \frac{\sigma_e^2}{n_{ri}}
$$

Where:

- $\frac{\sigma_f^2}{F_i}$ = form effects, resulting from multiple versions of the assessment with varying levels of difficulty.
- $F_i$ = number of forms administered at school $i$.
- $\frac{\sigma_{sf}^2}{F_i}$ = school by form interactions, due to alignment between form and curriculum.
- $\frac{\sigma_p^2}{n_{ri}} \cdot \frac{N_{ri} - n_{ri}}{N_{ri}}$ = error generated through pupil sampling from finite population.
- $N_{ri}$ = number of pupils in school $i$.
- $n_{ri}$ = number of pupils tested, considering all forms, in school $i$.
- $\sigma^2 = \text{variance in pupil observed scores (pooled within school/within form)}$.
- $R = \text{proportion of observed score variance relative to true score}$.
- $\sigma_p^2 = \sigma^2 \cdot R = \text{variance in pupil true scores (pooled within school)}$.
- $\sigma_e^2 = \sigma^2 \cdot (1 - R) = \text{variance in pupil error scores (pooled within school/within form)}$.

Sampling a group of students from a finite population can generate extreme shifts in ability, dependent upon the sample selected. The same can be true when selecting a form.
or version of an assessment from a group with varying levels of difficulty. While all
types of error contribute to the variability of assessment results, the primary concern for
this study will be sampling error resulting from cross-sectional analysis.

Sampling error resulting from cross sectional analysis, or testing a different group
of students each year, is a major concern when reporting percentage of students reaching
standards (Arce-Ferrer, Frisbie, and Kolen, 2002; Hill & DePascale (b), 2003; Kane &
Staiger, 2002; Linn & Haug, 2002; Miller, 2003; Yen, 1997). One could argue that
evaluating a different group of students each year is analogous to evaluating a new school
each year, thereby making year to year comparisons invalid. However, if viewed in
terms of the infinite population model, a group of students enrolled in a particular grade
in a particular year in a particular school could represent a random sample from the
groups that have enrolled in previous years as well as the infinite number of groups likely
to enroll in future years (Arce-Ferrer, Frisbie, and Kolen, 2002; Cronbach, Linn,
Brennan, & Haertel, 1997; Yen, 1997). The primary goals of NCLB are to evaluate the
quality of the school’s academic program, not its effectiveness with a particular group of
students. Assuming school factors remain relatively constant from year-to-year,
proportions of proficient grade five students should also remain relatively constant.

Within the infinite population model, sample error resulting from testing a finite
group of students from an infinite population must be considered, even if all students
enrolled during that particular test administration are assessed. In any given year, a
school will assess a sample of students from their infinite number of past, present, and
future students. The performance of this group will vary, depending on the abilities of
the group reaching the appropriate age during that particular testing cycle. In an attempt
to differentiate change due to improvement from change due to sampling error, Arce-Ferrer, Frisbie, and Kolen (2002) have concluded school level sampling error is large enough to interfere with annual change estimates. They discovered about two-thirds of the variability of estimates of change in proportions was due to sampling error with intervention effects, systematic errors, measurement errors, and equating errors accounting for the additional one-third. Kane & Staiger (2001) estimated an average 65 student fourth grade reading or math class would have an error interval extending from the 25th to the 75th percentile, approximately. Considering the average elementary school enrollment is 65 students, most would be too small to provide accurate PAC results, creating a likely environment for inaccurate NCLB reporting (Kane & Staiger, 2001; Yen, 1997).

To compensate for error associated with testing a sample of students from an infinite population, Nevada uses a 95% upper-tail confidence interval. According to the Nevada Adequate Yearly Progress Technical Manual (LaMarca, 2006), the 95% upper-tail confidence interval is defined as:

\[ C.I. = 1.645 \sqrt{\frac{P(1-P)}{n}} \]

Where:

\( P \) = proportion of proficient students

\( n \) = number of students assessed

\( \sqrt{\frac{P(1-P)}{n}} = \) sample error

1.645 = z-score consistent with 95% upper-tail confidence interval
If the PAC rate plus confidence interval is greater than or equal to the target rate of proficiency, the group is said to have made Adequate Yearly Progress (Porter, Linn, & Trimble, 2005). For example, a sample of 45 students with a 40% rate of proficiency could expect to generate a 12.01% margin of error. This corresponds to a 95% upper-tailed certainty that the true rate of proficiency could fall anywhere between 40.00% and 52.01%. Also noteworthy, if the increase in rate of proficiency from the previous year is less than 12.01%, it would be difficult, if not impossible, to distinguish actual growth from sample error (Arce-Ferrer, Frisbie, and Kolen, 2002).

The actual PAC rate, or observed score, is simply an estimate of the schools true proportion of proficient students. The true proportion, resulting from the sum of the confidence interval and the observed score, is necessary to satisfy the requirement for statistically reliable and valid results and should always be used when making inferences about a school’s performance (Coladarci, 2003). An unfortunate consequence of using true score is the possibility a school could achieve AYP when they should not have, resulting in the loss of resources and options afforded to schools designated as in need of improvement (Hill & DePascale (b), 2003).

As mentioned earlier, sample error, and the associated confidence interval, is inversely proportional to the number of students tested, meaning as subgroup size decreases margin of error increases. An interesting example of not recognizing sample size and its influence on error comes from Kane & Staiger’s “Volatility in School Test Scores: Implications for Test Based Accountability Systems” (2001):

When the 1998-1999 Massachusetts Comprehensive Assessment System test scores were released in November of 1999, the Provincetown District showed the
greatest improvement over the previous year. The Boston Globe published an extensive story describing the various ways in which Provincetown had changed educational strategies between 1998 and 1999, interviewing the high school principal and several teachers. As it turned out, they had changed a few policies at the school - decisions that seem to have been validated by the improvement in performance. One had to dig a bit deeper to note that the Provincetown high school had only 26 students taking the test in 10th grade (pg. 236).

Subgroup Size

When determining minimum subgroup size for reporting purposes, NCLB states subgroup data “shall not be required in a case in which the number of students in a category is insufficient to yield statistically reliable information (Public Law 107-110).” This open-ended approach to minimum subgroup size must be carefully scrutinized as both small and large numbers have consequences. Smaller minimum n-counts result in larger percentages of schools reporting results with a greater number of subgroups reporting per school (Porter, Linn, & Tremble, 2005). Small minimum n-count could also result in statistically unreliable results due to random error fluctuations. Large minimum n-counts will reduce the number of reported schools and subgroups, eliminating the benefits of disaggregated reporting while transferring much of the burden of accountability to large schools (Linn, 2003; Linn, Baker, & Herman, 2002; Marion, et al., 2002).

Hill & DePascale (2002) concluded schools with as few as 20 students were correctly classified approximately 85% of the time. When minimum n-count was

28
increased to 50 they found the accuracy of classifications increased to around 90%.

While many states have selected a minimum n-count of 30, largely due to 30 being the point in the \( z \) or \( t \) statistical tables where values level off (Marion, et al., 2002), Nevada has selected 25 as their minimum n-count for AYP reporting purposes (LaMarca, 2006). While states have selected a variety of minimum n-counts for reporting purposes, it appears a reasonably reliable system of classifications is possible when working with relatively small schools and subgroups.

A final concern is the impact of minimum n-count on highly diverse schools. If a state requires schools to report the performance of small subgroups, it is possible for a highly diverse school to report all subgroups for English/Language Arts and Math. According to Linn, Baker, & Herman (2002), if the school is serving the academic needs of all its subgroups equally, and has a 70% probability of reaching the target rate of proficiency, with a single subgroup reporting the school has a 70% chance of making AYP. But, due to the independent nature of subgroup fluctuations, a school with two reporting subgroups will have a 49% chance of making AYP \((.7 \times .7 = .49)\). This pattern continues until, with the maximum 16 subgroups reporting, the school will have a 0.33% probability of making AYP \((.7^{16})\), meaning 33 of every 1000 schools in this particular situation would make AYP with 967 failing. In other words, the more subgroups a school is forced to report, the less likely the school will make AYP.

Mathematically speaking, NCLB is well designed and effectively provides a cookie-cutter approach to academic evaluation. Progressive sanctions have been included that gives ample time for schools to adjust programs with baseline proficiency rates established that insure 80% of all schools make school-level AYP the first year. Periodic
increases in target rate of proficiency have been staggered to allow schools time for
adjustment with a safe-harbor provision included to reward schools not achieving the
targeted rate of proficiency but demonstrating sufficient positive growth. And finally,
reliable and valid assessments have been provided, reporting the performance of
subgroups having 25 or more students, with compensations included for measurement
error. Unfortunately, not considered in NCLB were accommodations for schools the law
was specifically designed to target; schools housing larger than average numbers of high-
needs, low-performing students.

Subgroup Performance and Diverse Schools

There are a plethora of studies devoted to the academic challenges of special
needs and racially diverse schools (Koretz & Hamilton, 2000; Bankston & Caldas, 1998;
requirements for meeting AYP pose the greatest challenges to high poverty schools
which enroll a large percentage of students that have traditionally scored poorly on
standardized achievement tests.” Taking into consideration the high correlation between
race and special populations, minority subgroups are more likely to belong to
economically depressed and limited English subgroups, thereby increasing their
propensity to not meet requirements through NCLB (Kane, Staiger & Gippert, 2001; Kim
& Sunderman, 2005; Orfield, 1996; Orfield & Lee, 2005). This would imply schools that
enroll large concentrations of poor, minority students would be greatly disadvantaged
when attempting to meet the requirements of NCLB.
Requiring schools to report the performance of subgroups with 25 or more students may be counterproductive to the goals of NCLB. The so-called subgroup rules have resulted in a decrease in resources and an increase in sanctions to racially diverse schools with no measurable impact to minority performance (Kane & Staiger, 2003). While the increase in attention to minority group performance was necessary and advantageous to previously ignored populations, subgroup reporting has singled out for sanctions and excluded from any reward system schools with large diverse populations. It has been predicted that elevated AYP failure rates will be experienced in high poverty schools, as measured through eligibility for free or reduced lunches, and schools with large minority populations (Chubb, Linn, Haycock, & Wiener, 2005; Kim & Sunderman, 2005). By comparison, predominately white schools with homogeneous enrollments of 250 or less will rarely fail to achieve AYP due to subgroup sizes smaller than reporting minimums (Rose, 2004). In a recent report on segregation, Orfield and Lee (2005) found:

White students are the least likely subgroup to attend multiracial schools, with black and Hispanic students most likely to attend schools with a majority racial group as their own. The average black student will attend a school with approximately 12.5% Hispanic, an average Hispanic will attend a school with a similar proportion of black students, and both racial groups, on average, attend schools that are 30% or less white. In contrast, the average white student will be enrolled in a school where nearly 80% of the students are also white. Asians have been found to be the least isolated within their own racial group, attending schools that are, on average, 45% white, 12% black, and 20% Hispanic (pg. 12).
Segregation by race has systematically been linked to other forms of segregation, such as segregation by socioeconomic status, residential location, and language. Orfield and Lee (2005) find insulting the implication an all Black or Hispanic school is somehow academically inferior. They have discovered a strong correlation exists between percent poor and percent minority, with academic differences associated with socioeconomic status, not race.

Another contributor to the achievement gap is health issues common to poor and low-income families. Elevated cases of lead poisoning, vision and hearing problems, cytomegalic inclusion disease, asthma, psychosocial and psychosomatic problems, and iron deficiency anemia are found in poor children (Egbuonu & Starfield, 1982; Starfield, 1982). In 2002, 20.1% of children living in poverty had no health insurance (Mills & Bhandari, 2003), with poor children 75% more likely to be admitted into a hospital during any given year with the average total hospital stay, or missed school days, four times greater than their more affluent classmates (Starfield, 2002). Fetal Alcohol Syndrome, or FAS, and a variety of related prenatal behaviors such as smoking, poor nutrition, poor health, increased stress, and use of drugs have been directly linked to the low socioeconomic status of the mother (Abel, 1994). These risky prenatal behaviors often result in premature or low-birth weight babies with a variety of cognitive limitations such as low IQ, learning disabilities, and attention disorders (Hack, Kline, & Taylor, 1995). While race seems to be the popular metric for evaluating differences in the behavior of children, socioeconomic status appears to be the leading factor. Abel (1995), in his study of FAS found “Although race and SES are confounded in the U. S.
studies, an examination of U.S. and European studies suggests that the major factor associated with FAS is low SES rather than racial background (437).

Rothstein (2004) hypothesizes children from poverty, even in the best of schools, will achieve less than middle class students. Parents from different educational backgrounds raise children differently, with more educated parents reading to and encouraging their children to read (Bianchi & Robinson, 1997; Hofferth & Sandburg, 2001) and young children of college-educated parents exposed to more books in the home than children of less educated parents (Denton & Germino-Hauskens, 2000). Rothstein (2004) describes the current gaps in achievement and its relationship to poverty as:

Some low-income children are naturally quick learners, take to school well, and respond so well to high expectations that after a few years of school they read better than typical middle-class children. Some middle-class children get no support for learning from troubled families, and some low-income parents organize life around a dream of college. But, on average, a typical middle-class child who began to read at home will have higher lifetime achievement than a typical low-income child who was taught only in school, even if each benefits from good curriculum, effective teaching, and high expectations (19).

Pallas, Natriello, and McDill (1989) view education as a combination of experiences, collected through formal education, family interactions, and community involvement. They believe any student exposed to inadequate or inappropriate experiences from any of the three aforementioned domains would be considered educationally disadvantaged. In an examination of educational disadvantage encountered
by black and Hispanic students, it was found that the educational disadvantage was not associated with membership in a minority group but rather with living below the poverty line (Jimerson, Egeland, & Teo, 1999; Pallas, Natriello, & McDill, 1989; Schwartz, Yen, & Schafer, 2001). Entwisle and Alexander (1996) reinforced this theory when they found:

- School poverty level overshadows racial profile when explaining majority/minority achievement differences.
- While school is in session, poor students and more affluent students performed at nearly the same level.
- When school is not in session, such as summer break, students in poverty suffer academic loss, whereas students not in poverty experience academic gain.

This does not imply all poor, minority children are educationally disadvantaged. Nonetheless, on average these characteristics, along with living in a single parent home, having a poorly educated mother, and having a limited English-speaking background have been associated with lower levels of academic achievement (Pallas, Natriello, & McDill, 1989). Aggregated results that do not recognize and adjust for differences in student backgrounds associated with poverty are unlikely to stimulate adequate performance improvements and will possibly distort the school’s academic effectiveness (Orfield & Lee, 2005; Sicoly, 2002; Stone & Lane, 2003).

**Students with Limited English Proficiency**

Large scale assessments have traditionally excluded English language learners, largely due to confounding variables associated with language proficiency and academic
achievement (Abedi, Hofstetter, & Lord, 2004). Low educational attainment for limited English proficient students has been attributed to racial and ethnic segregation in poor, underfunded urban schools (Schmid, 2001). Rumberger and Gándara hypothesize the academic achievement of limited English proficient (LEP) students lag that of English background students based upon seven inequitable conditions:

1. Limited access to appropriately trained teachers
2. Inadequate professional development for teachers of LEP students
3. Limited or no access to appropriate assessment necessary to gauge learning needs or progress
4. Inadequate instructional time to address needs and accomplish goals
5. Limited access to appropriate instructional materials and curriculum
6. Inadequate facilities and instructional environments
7. High incidents of segregation, placing them at risk for academic failure

While all are equally important, it is the intention of this study to focus on the lack of appropriate assessment materials necessary to evaluate needs and growth. Rumberger & Gándara (2004) found using an inappropriate assessment tool can have serious negative results, regardless of the outcome. Their findings suggest positive change interpreted as a gain in subject matter may represent nothing more than an increased level of English proficiency whereas low performance can have the opposite effect, prompting remedial studies or even special education interventions when the student has mastered the subject matter but cannot express the necessary skills in English. Extensive research on the assessment of LEP students has found performance gaps between LEP and non-LEP
students can be attributed to a lack of English language proficiency, not content knowledge (Abedi, 2004).

Nevada students are classified limited English proficient using a combination of home language survey and annual assessment of English proficiency (U. S. Department of Education, 2005). When participating in state mandated testing, such as assessments used for determining AYP, LEP students can be given certain test accommodations designed to level the playing field without providing unfair advantages (Abedi, Hofstetter, & Lord, 2004). Currently, Nevada provides a variety of testing accommodations to LEP students, which are intended to improve access to the assessment and its content while maintaining test validity and comparability of scores (NPEP Guidelines, 2006). Available accommodations include (Appendix A):

1) Accommodations in the test setting such as:
   a) individual administration
   b) small group administration
   c) administration in an alternative setting
   d) testing in a study carrel or reasonable substitute.

2) Accommodations in test administration:
   a) having a specific individual administer (i.e., ESL/ELL teacher)
   b) use of a bilingual dictionary or electronic translator (single word-at-a-time translation)
   c) have questions answered regarding specific testing procedures
   d) directions read aloud at the beginning of the test, word for word, in the student's native language
e) reread aloud the directions at the beginning of the test, word for word, in English
f) read the mathematics test(s) word for word, text only, in English
g) read the science test(s) word for word, text only, in English
h) read the writing prompt word for word, in English

An effective accommodation is one that can be administered to both LEP and non-LEP students without threat to test validity (Abedi, Hofstetter, & Lord, 2004; American Educational Research Association, 2004). Nevada limits the accommodations available for non-English speaking students to options that will assist in understanding expectations without altering the construct being measured.

Unfortunately, testing non-English speaking students with English-only assessments introduces construct irrelevant components, reflecting in part current levels of English proficiency instead of abilities (American Educational Research Association, 2004). Standard 9.1 of the Standards for Educational and Psychological Testing (American Educational Research Association, 2004) states “Testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences,” meaning any inferences introduced concerning a non-English speaking students abilities are inappropriate when using English language assessments. While a simple translation into the student’s native language seems appropriate, this approach would not produce an assessment equivalent in content, difficulty, reliability, and validity (Abedi, Hofstetter, & Lord, 2004; American Educational Research Association, 2004). Translating words across languages, even if they appear similar, can take on different meanings, especially if a variety of dialects are involved. It has been suggested that assessments using languages other than English
should be administered only to those students receiving instruction in that language and are familiar with the content terminology of that language (Abedi, Hofstetter, & Lord, 2004).

Many ethnic and special populations, as evaluated through NCLB, have consistently demonstrated a lower overall performance (Koretz & Hamilton, 2000; Bankston & Caldas, 1998; Lee, 2002; Abedi, 2004; Ogbu, 1994). A variety of studies identify possible causes for the achievement gap, such as family structure (Bankston & Caldas, 1998), inclusion and accommodations (Koretz & Hamilton, 2000), test validity and subgroup stability (Abedi, 2004), simple racial and ethnic achievement gaps (Lee, 2002), and racial stratification (Ogbu, 1094). LaMarca (2006), in a summary of the challenges of meeting the requirements of NCLB, stated:

The one-size fits all approach required by the NCLB legislation and the ability to produce meaningful test scores present a quandary for states. The act prevents a consideration of a variety of social factors that contribute to test score variance. But we have strong evidence that these factors do in fact affect our interpretations of performance..... Given the constraints of the legislation, states are left with understanding the effects of sociocultural factors in a post hoc fashion as they evaluate the impact of their programs.

**Summary**

NCLB requires states to develop, under very specific conditions, a valid and reliable system of measurement that meets certain requirements while neglecting others. One such provision is a system of assessments that holds all schools to the same
academic criteria, regardless of circumstance, while at the same time acknowledging an achievement gap exists between certain subpopulations of students. Nevada has developed such a program, which effectively fulfills the requirements of NCLB, and like NCLB, neglects to recognize the unique challenges for those schools housing larger than average populations of lower-performing students. Included in the Nevada Plan was a valid and reliable assessment system, baseline achievement levels that fairly accommodated Nevada schools, compensation for measurement error, minimum n-counts for reporting that appear to closely maximize subgroup performance, and provisions that allow recognition for sufficient annual growth. Missing from the Nevada Plan, due to federal restrictions, was any type of flexibility from sanctions for the large, diverse, lower achieving schools. As the literature has demonstrated, schools must annually maintain a higher level of achievement while dealing with issues of cohort instability, making the task near impossible. Also problematic is the number of reported subgroups. As the diversity of the school increases, the number of reported subgroups also increases, reducing the probability of making AYP.

Historically, when discussing the achievement gap, much attention has been placed on social difficulties of certain subpopulations, which may or may not contribute. It is not the intention of this study to examine social issues associated with racial subgroups or special populations of students. Regardless of the contributing factors, it will suffice to hypothesize a difference in subgroup performance exists. While a correlation between low-performing schools and minority, special population students appear to be implied, this study in no way supports causation. It is simply the intention
of this study to examine the academic performance of schools within homogeneous clusters with no considerations of achievement across clusters.
CHAPTER 3

RESEARCH DESIGN AND METHODOLOGY

Introduction

The No Child Left Behind Act of 2002 (NCLB; Public Law No. 107-110, 115 Stat. 1425, 2002) was designed to guarantee all students receive a fair, equitable, and high-quality education. One of the major provisions tied to NCLB is a measurable Adequate Yearly Progress (AYP) using state-developed content standards based assessments in both English/Language Arts and mathematics. State, district and school AYP measurements are reported for all students, as well as subgroups of students defined by major ethnic groups (American Indian/Alaskan Native, Asian/Pacific Islander, Hispanic, African American, and White), students with Individual Education Plans (IEP), students with Limited English Proficiency (LEP), and students eligible for Free or Reduced Lunches (FRL). In addition to annual reports of percentage of proficient students, states, districts and schools must also report the percentage of students participating in standards based testing and the school wide average daily attendance or graduation rate.

A key element in NCLB is the fair and equitable treatment of all schools, regardless of local circumstances. The Consolidated State Application Accountability Workbook, published by the U.S. Department of Education (U.S. Department of Education).
Education, 2005) provides guidance for designing an accountability system and suggests guidelines as determined by NCLB. As mandated through NCLB, all state accountability workbooks must include:

- An accountability system which includes all schools
- An accountability system which includes all students
- An accountability system that includes all major subgroups
- An accountability system that properly includes mobile students
- An accountability system holding all schools to the same criteria
- An accountability system that includes rewards and sanctions

All states, districts, and schools will be evaluated using the same criteria, regardless of unique conditions. For example, schools having a minority English speaking population will be judged using the same criteria as a school with a majority of English-speaking students. Although prior research has demonstrated the demographic profile of a school will have an influence on student performance, NCLB disregards all unique circumstances and requires schools be evaluated using a standardized system of accountability.

Statement of the Problem

NCLB mandates schools be evaluated using a standardized methodology without consideration for unique factors such as larger than average minority or special needs populations. While designed to eliminate the achievement gap between high and low performing students, NCLB, in its current design, singles out for sanctions and excludes from rewards schools serving racially diverse, special needs student populations (Kane,
Staiger & Geppert, 2001; Kim & Sunderman, 2005). Considering the high stakes associated with failing to meet AYP, is the statute mandating schools be evaluated using a standardized approach without first considering academic achievement relative to schools with similar demographic profiles sound federal policy?

**Purpose of the Study**

It is the purpose of this study to examine the efficacy of using Nevada standards based CRT results to determine a school's overall academic performance and annual progress without consideration for unique demographic characteristics. It is the belief of this researcher that a school's dominant student population will have a significant influence on academic performance, which if not considered could result in grave consequences with respect to NCLB. Using the TwoStep cluster analysis procedure, schools will be classified and homogeneous clusters of schools formed using the criteria dominant student populations. Determinations will then be made concerning the statistically significant differences in mean reading and mean math CRT scale scores for schools contained within discrete homogeneous clusters. This study will determine if academic performance is influenced by a school's demographic profile, and if so, are schools sharing similar demographic signatures performing as expected.

**Research Questions**

The research questions for this study are:

1) Which demographic variables, as recognized through NCLB, generate a unique and homogeneous school cluster of five or more schools?
a) Demographic variables generate unique and homogeneous clusters when examining the American Indian/Alaskan Native subpopulation.

b) Demographic variables generate unique and homogeneous clusters when examining the Asian subpopulation.

c) Demographic variables generate unique and homogeneous clusters when examining the Hispanic subpopulation.

d) Demographic variables generate unique and homogeneous clusters when examining the African American subpopulation.

c) Demographic variables generate unique and homogeneous clusters when examining the White subpopulation.

f) Demographic variables generate unique and homogeneous clusters when examining the IEP subpopulation.

g) Demographic variables generate unique and homogeneous clusters when examining the LEP subpopulation.

h) Demographic variables generate unique and homogeneous clusters when examining the FRL subpopulation.

2) Does a statistically significant difference in mean reading CRT scale score exist within homogeneous clusters of five or more schools?

a) If the subgroup American Indian/Alaskan Native generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.
b) If the subgroup Asian generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

c) If the subgroup Hispanic generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

d) If the subgroup African American generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

e) If the subgroup White generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

f) If the subgroup IEP generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

g) If the subgroup LEP generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

h) If the subgroup FRL generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean reading CRT scale score.

3) Does a statistically significant difference in mean math CRT scale score exist within homogeneous clusters of five or more schools?
a) If the subgroup American Indian/Alaskan Native generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

b) If the subgroup Asian generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

c) If the subgroup Hispanic generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

d) If the subgroup African American generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

e) If the subgroup White generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

f) If the subgroup IEP generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

g) If the subgroup LEP generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.
h) If the subgroup FRL generated a unique and homogeneous cluster of five or more schools, certain schools will demonstrate a statistically significant difference in mean math CRT scale score.

Research Design

Cluster analysis is a standard term used to describe the group classification or "clustering" of items sharing similar attributes. More specifically, cluster analysis is a multivariate statistical procedure used to reorganize data sets into relatively homogeneous groups (Aldenderfer & Blashfield, 1984; Bartholomew et. al., 2002; Anderberg, 1973; Hair & Black, 1998). A simple example would be the consolidation of test scores, measured in percentage correct, into five clusters, one for each grade “A”, “B”, “C”, “D”, and “F”. Using test score as the data points, with percentage correct the reference point, letter grades could eventually be used to replace the numeric reference points (Faber, 1994). This optimization process would continue until no data points change clusters (Aldenderfer & Blashfield, 1984; Bartholomew et.al., 2002; Anderberg, 1973; Hair & Black, 1998).

TwoStep cluster analysis is a scalable algorithm designed to complete the classification process in two steps. In a single data pass the TwoStep procedure pre-clustering all cases into many small sub-clusters and then, using an agglomerative hierarchical procedure clusters the small sub-clusters into a user defined number of final clusters. While primarily designed to handle large data sets, the attraction of the TwoStep procedure in this study is its ability to handle categorical variables such as school name (SPSS, 2001; SPSS, 2006).
The pre-cluster process used a sequential approach of scanning each data point to determine if the current data point should be merged into an already formed cluster or form a new cluster. TwoStep uses a modified cluster feature (CF) tree with levels of nodes and leaf nodes. The SPSS TwoStep Cluster Component Technical Paper (2001) and SPSS 15.0 Algorithms (2005) define the TwoStep procedure as:

Step 1: The pre-cluster step

Records are scanned one by one and determinations made, based upon the distance criterion, if the record should merge with a previously formed cluster or form a new cluster. The CF tree has levels of nodes with each node containing a number of entries with each leaf entry representing a sub-cluster. Non-leaf nodes and their entries quickly guide new records into a correct leaf node. The SPSS default CF tree has, as a maximum, three levels of nodes with eight entries per mode, allowing at most $8^3$ or 512 leaf entries or subclusters.

Each entry is differentiated by the CF that contains the entries number of records, the mean and variance of each continuous variable, and counts for each category of each categorical variable. Upon reaching a leaf node, the entry finds the closest leaf entry and, if within the threshold value, is absorbed by the leaf node and the CF is updated. If not within the acceptable threshold level, the entry will create its own leaf entry in the leaf node. If space is not available in the leaf node to create a new leaf entry the leaf node splits into two separate leaf nodes and redistributes the entries based upon the closest criteria using the farthest pairs as seeds. Should the CF tree exceed the maximum number of levels, the CF tree rebuilds itself with an increased threshold distance criterion. The rebuilt CF tree

48

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
is then smaller, allowing space for new input records. This process is continued through one complete data pass. It is suggested the data be ordered randomly to avoid bias resulting from sequential patterns.

Step 2: The cluster step

The cluster step creates a user-defined number of clusters from the sub-clusters formed in the first step. Using an agglomerative hierarchical clustering method, all sub-clusters are compared with the pair of sub-clusters demonstrating the smallest distance merged into a single cluster. Once merged, the new sets of clusters/sub-clusters are again compared with the pair demonstrating the smallest distance again merged. This procedure continues until all clusters have been merged.

Because the variable school name was categorical it was necessary to use as the distance measurement for this study the log-likelihood criterion, a probability based distance formula that assumes normal distributions for continuous variables, multinomial distributions for categorical variables, and independence for all variables and cases. When combining clusters the distance between two clusters is related to the decrease in log-likelihood with the distance between clusters \( j \) and \( s \) defined as (SPSS, 2005):

\[
d(j, s) = \xi_j + \xi_s - \xi_{<j,s>}
\]

(Eq. 1)

Where:
\[ \xi_v = -N_v \left( \frac{1}{2} \sum_{k=1}^{K^A} \log \left( \hat{\sigma}_k^2 + \hat{\sigma}_{vk}^2 \right) + \sum_{k=1}^{K^B} \hat{E}_{vk} \right) \tag{Eq. 2} \]

\[ \hat{E}_{vk} = -\sum_{i=1}^{L_k} \frac{N_{ski}}{N_v} \log \frac{N_{ski}}{N_v} \tag{Eq. 3} \]

And:

\[ K^A = \text{Total number of continuous variables used in the procedure.} \]
\[ K^B = \text{Total number of categorical variables used in the procedure.} \]
\[ L_k = \text{Number of categories for the } k^{th} \text{ categorical variable.} \]
\[ R_k = \text{The range of the } k^{th} \text{ continuous variable.} \]
\[ N = \text{Number of data records in total.} \]
\[ N_k = \text{Number of data records in cluster } k. \]
\[ \hat{\sigma}_k^2 = \text{The estimated variance of the } k^{th} \text{ continuous variable in whole data.} \]
\[ \hat{\sigma}_{jk}^2 = \text{The estimated variance of the } k^{th} \text{ continuous variable in cluster } j. \]
\[ N_{ski} = \text{Number of data records in cluster } j \text{ whose } k^{th} \text{ categorical variable takes the } i^{th} \text{ category.} \]
\[ d(j,s) = \text{Distance between clusters } j \text{ and } s. \]
\[ <j,s> = \text{Index that represents the cluster formed by combining clusters } j \text{ and } s. \]

Ignoring \( \hat{\sigma}_k^2 \) in equation 2 would result in the exact log-likelihood decrease between clusters \( j \) and \( s \) after the two clusters are combined. The \( \hat{\sigma}_k^2 \) term was added to avoid problem caused when \( \hat{\sigma}_{vk}^2 = 0 \), which would result in an undefined natural logarithm.
The initial TwoStep cluster procedure generated eight unique clusters, corresponding to the eight standardized variables quantifying proportion of school population (American Indian, Asian, Hispanic, African American, White, IEP, LEP, FRL), with the number of clusters equal to the number of variables. If \( n \) dependent variables did not generate \( n \) unique homogeneous subsets, or when dependent variables demonstrated strong multicollinearity, the dependent variables were reduced by \( n \) to eliminate ineffective or irrelevant variables and the TwoStep cluster procedure repeated with \( n \)-less clusters (Hair & Black, 1998).

When \( n \) homogeneous clusters were generated, each statistically significant with respect to a single unique variable, the One-Way Analysis of Variance (ANOVA) procedure was performed on individual cluster mean reading and mean math CRT scale scores. The ANOVA produced a one-way analysis of variance for the quantitative dependent variable school mean reading or mean math CRT scale score by independent variable homogeneous cluster. The ANOVA tested the null hypothesis there were no statistically significant differences in mean reading or mean math CRT scale scores for schools located within each homogeneous cluster. When a statistically significant difference in the mean reading or mean math CRT scale score was found, and within cluster schools demonstrated equal variances, Tukey’s post-hoc test was used to identify which schools performance on the reading or math CRT was significantly different from the others. In such cases where variances were unequal, Tamhane’s T2 post hoc procedure was used.
Data and Data Sources

The data collected for this study was the spring, 2006 grades three through five Nevada Criterion Referenced Test (CRT) results. The tests were administered and data collected from all Clark County School District grades three, four, and five students enrolled during the spring, 2006 CRT test administration. Elementary schools operating on the traditional 9-month calendar administered the CRT during the test window beginning March 1, 2006 and ending April 17, 2006. To insure all students received 120 ± 10 instructional days prior to CRT testing, elementary schools operating on the 12-month, or year-round, schedule administered the CRT to track 1 students during the traditional 9-month testing window with tracks 2-5 administering the test during the testing window beginning April 17, 2006 and ending May 5, 2006. Schools could administer the tests at any time during the scheduled testing window with answer documents returned to the Clark County School District testing department on or before the final day of the scheduled testing window.

Inclusion in this study was contingent upon enrollment for the full academic year (yis = 1) as well as active participation on both the reading and math CRT. The data collected included student-level reading and math scale scores, gender, ethnic group, and membership in the IEP, LEP, or FRL subgroup. Ethnicity was identified by parent or guardian and entered into the student data system at time of enrollment with IEP classification dependent upon parent and/or teacher referral and cognitive assessment results. All students identified as requiring an individual education plan, with the exception of students identified as gifted or talented, were included in the IEP subgroup. Identification as limited English proficient resulted from non-English being the primary
language spoken in the home with participation in the Language Assessment Scale and a combined score less than 241 (on a scale of 1-300). Inclusion in the FRL subgroup was based upon household income with eligibility determined through completion of the Family Application of Meal Benefits. This study included 163 Clark County School District Elementary schools with 22,150 students actively participating in the spring, 2006 administration.
CHAPTER 4

DATA ANALYSIS

Introduction

The purpose of this study was to examine the efficacy of using Nevada standards based CRT results to determine a school's overall academic performance and annual progress without first considering unique demographic characteristics. Using the TwoStep cluster analysis procedure, schools were classified based upon dominant student populations with determinations made concerning statistically significant differences in mean reading and mean math CRT scale scores for those schools contained within homogeneous clusters. This study was designed to determine if academic performance was in fact influenced by demographic profile, and if so, were schools sharing similar demographic signatures performing as expected relative to their unique student population.

This study began by standardizing the proportions of ethnic and special population students enrolled at each individual school site. The TwoStep cluster analysis procedure was then used to define eight unique clusters, corresponding to the eight standardized subgroup variables (American Indian, Asian, Hispanic, African American, White, IEP, LEP, FRL) with the number of cluster centroids equal to the number of subgroup variables. Subgroup level of significance with respect to discrete identified
cluster was determined using the studentized t-procedure with Bonferroni adjustments applied. When any identified subgroup did not generate a unique homogeneous cluster or instances of strong multicollinearity were found to exist between subgroups, those subgroups were determined ineffective or irrelevant, excluded, and the cluster procedure repeated.

Once \( k \) unique and homogeneous clusters were identified, the One-Way ANOVA procedure was performed on individual cluster mean reading and mean math CRT scale scores to test the null hypothesis no statistically significant difference in mean reading or mean math CRT scale scores existed. In cases where it was determined a statistically significant difference in the mean reading or math CRT scale score did exist, appropriate post-hoc tests were used to identify which school’s performance on the reading or math CRT was significantly different from the others.

**Subgroup Standardized Proportions**

Data analysis began with a breakdown of assessed student population, using subgroup n-counts and percentages, with standardizing proportions relative to district-wide subgroup percentages. Rural as well as special schools were eliminated from the data set to avoid bias resulting from small size and non-urban school settings. Eliminated were Child Haven ES (Special School, \( n=1 \)), Child Haven ES Detention (Educational Services, \( n=1 \)), Miley Achievement Center (Special School, \( n=4 \)), Variety ES (Special School, \( n=8 \)), Bennett ES (Laughlin, Nv., \( n=117 \)), Blue Diamond ES (Blue Diamond, Nv., \( n=13 \)), Grant Bowler ES (Logandale, Nv., \( n=351 \)), Joseph Bowler ES (Bunkerville, Nv., \( n=227 \)), Goodsprings ES (Goodsprings, Nv., \( n=4 \)), Indian Springs ES (Indian
Springs, Nv., n=50), Martha P. King ES (Boulder City, Nv., n=468), Lundy ES (Mt Charleston, Nv., n=3), Perkins ES (Moapa Valley, Nv., n=85), Reid ES (Searchlight, Nv., n=17), Sandy Valley (Sandy Valley, Nv., n=77), and Virgin Valley ES (Mesquite, Nv., n=284). The remaining 175 Clark County School District elementary schools were retained in the data set for analysis. Student inclusions were dependent upon enrollment for the full academic year (YIS = 1), not classified as new in country (NIC = 0), and actively participating in both reading and math portions of the CRT. Assessed subgroup n-counts with assessed percentages and district wide percentages are summarized in table 1 with individual school n-counts and percentages found in Appendix B. The large difference between assessed and district LEP percentages results from district reporting only current limited English proficient students while NDE includes current as well as former LEP (FLEP) when reporting academic performance.

Table 1. Subgroup n-counts and percentages

<table>
<thead>
<tr>
<th></th>
<th>Assessed N-Count</th>
<th>Assessed Percentage</th>
<th>District Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>District</td>
<td>58894</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>American Indian</td>
<td>473</td>
<td>0.80%</td>
<td>0.84%</td>
</tr>
<tr>
<td>Asian</td>
<td>5,210</td>
<td>8.85%</td>
<td>8.45%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>22,772</td>
<td>38.67%</td>
<td>36.84%</td>
</tr>
<tr>
<td>African American</td>
<td>7,779</td>
<td>13.21%</td>
<td>14.42%</td>
</tr>
<tr>
<td>White</td>
<td>22,660</td>
<td>38.48%</td>
<td>39.45%</td>
</tr>
<tr>
<td>IEP</td>
<td>6,234</td>
<td>10.59%</td>
<td>10.79%</td>
</tr>
<tr>
<td>LEP</td>
<td>16,372</td>
<td>27.80%</td>
<td>17.28%</td>
</tr>
<tr>
<td>FRL</td>
<td>28,601</td>
<td>48.56%</td>
<td>45.55%</td>
</tr>
</tbody>
</table>
When comparing within and among subgroup size, the use of percentages may create artificial differences. For example, the American Indian subgroup represents an insignificant percentage when compared across subgroups but may represent a significant percentage when compared to the percentage of American Indian students across schools. To eliminate these inappropriate comparisons, all subgroup percentages were standardized with a mean of 0 and standard deviation of 1 using the formula:

\[ Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \]

Where:

\( \hat{p} = \) school subgroup percentage

\( p = \) district subgroup percentage

\( n = \) school subgroup n-count or sample size

Once standardized, an American Indian z-score of 2.000 would represent 2 standard deviations above district percentage of American Indian populations across schools, allowing for comparisons with other subgroup z score within the school. Also, schools having subgroup n-counts of zero generated an undefined z-score (division by zero), creating a missing value for that school’s subgroup. In such cases where a data set has a limited number of missing values, it is appropriate to replace missing values with an educated guess or mean value calculated from available data (Tabachnick & Fidell, 1996). Therefore, the 24 schools void of an American Indian subgroup and the three schools with no Asian students were given z-scores of zero, representing the district percentage of American Indian (0.82%) and Asian (8.66%) students. Elementary school z-scores can be found in Appendix C.
Also eliminated from the data set were those schools not enrolling a significant number of students from any single subgroup. Using $\alpha = 0.05$, schools with $z < +1.96$ for all identified subgroups were removed from the data set. Eliminated from analysis were Adams ES, Brookman ES, Bruner ES, Bunker ES, French ES, Goldfarb ES, Gray ES, Guy ES, Hancock ES, Harris ES, Herr ES, McMillan ES, Sandy Miller ES, Simmons ES, Wasden ES, and Wolfe ES. The remaining 159 schools were retained in the data set for analysis due to enrolling at least one subgroup with a significant percentage above district enrollment ($z \geq +1.96$). And finally, in an attempt to avoid data set bias due to order all of the remaining 159 schools were assigned a random number and the data set sorted ascending relative to the randomly generated number.

Question 1: Which Demographic Variables, as Recognized Through NCLB, Generate Unique and Homogeneous School Clusters of Five or More Schools?

The initial TwoStep cluster analysis procedure generated eight unique clusters representing the eight demographic variables as recognized through NCLB. Table 2 catalogs initial cluster distribution, table 3 cluster centroids by subgroup variable, and Appendix D actual cluster membership by school. As table 2 illustrates, eight unique clusters were generated with a membership high of 30 schools for cluster 8 to a low of 12 schools assigned to clusters 5 and 6.
Table 2. Initial Cluster Distributions

<table>
<thead>
<tr>
<th>Cluster</th>
<th>N</th>
<th>% of Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
<td>11.95%</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>15.72%</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>11.32%</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>16.35%</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>7.55%</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>7.55%</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>10.69%</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>18.87%</td>
</tr>
<tr>
<td>Total</td>
<td>159</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Also provided are individual within-cluster variation charts quantifying overall and cluster mean z-scores, with 95% confidence limits around such means (see figure 1), and subgroup clusterwise importance charts with dashed vertical lines representing significance. When determining the clusterwise importance chart level of significance, Bonferroni adjustments were applied to control for type I error (see figure 2). Considering the purpose of this study was to identify clusters of schools serving subgroup populations significantly greater than district average, subgroup clusters must exceed the positive t statistic at $\alpha = 0.05$ (with Bonferroni adjustments applied) before statistical significance can be determined. Those variables not significant at the $\alpha = 0.05$ level of significance will not be displayed in the variable clusterwise importance plots.
Table 3. Initial Cluster Centroids

<table>
<thead>
<tr>
<th>Cluster Centroids</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AMERICAN INDIAN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.00</td>
<td>0.07</td>
<td>0.08</td>
<td>-0.01</td>
<td>0.17</td>
<td>0.05</td>
<td>-0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.04</td>
<td>0.15</td>
<td>0.08</td>
<td>0.03</td>
<td>0.27</td>
<td>0.09</td>
<td>0.03</td>
<td>0.12</td>
</tr>
<tr>
<td><strong>ASIAN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.29</td>
<td>0.76</td>
<td>-0.02</td>
<td>0.01</td>
<td>-0.38</td>
<td>5.35</td>
<td>-0.46</td>
<td>-0.40</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.78</td>
<td>1.10</td>
<td>0.53</td>
<td>0.59</td>
<td>0.21</td>
<td>2.79</td>
<td>0.24</td>
<td>0.38</td>
</tr>
<tr>
<td><strong>HISPANIC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-3.63</td>
<td>-2.82</td>
<td>-3.28</td>
<td>0.16</td>
<td>3.55</td>
<td>-3.07</td>
<td>16.03</td>
<td>7.92</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.34</td>
<td>0.80</td>
<td>0.51</td>
<td>2.63</td>
<td>4.31</td>
<td>0.63</td>
<td>2.88</td>
<td>3.26</td>
</tr>
<tr>
<td><strong>AFRICAN AMERICAN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-0.78</td>
<td>0.48</td>
<td>-0.67</td>
<td>5.24</td>
<td>0.45</td>
<td>-0.13</td>
<td>-0.68</td>
<td>0.30</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.47</td>
<td>1.69</td>
<td>0.31</td>
<td>7.52</td>
<td>1.17</td>
<td>0.48</td>
<td>0.32</td>
<td>1.48</td>
</tr>
<tr>
<td><strong>WHITE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>10.51</td>
<td>3.73</td>
<td>8.70</td>
<td>-1.21</td>
<td>-1.59</td>
<td>0.65</td>
<td>-2.86</td>
<td>-3.09</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>1.94</td>
<td>2.39</td>
<td>2.99</td>
<td>1.64</td>
<td>1.21</td>
<td>1.52</td>
<td>0.69</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>IEP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-0.31</td>
<td>0.11</td>
<td>0.53</td>
<td>0.12</td>
<td>1.44</td>
<td>-0.37</td>
<td>-0.26</td>
<td>-0.14</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.26</td>
<td>0.31</td>
<td>0.40</td>
<td>0.30</td>
<td>0.59</td>
<td>0.32</td>
<td>0.41</td>
<td>0.24</td>
</tr>
<tr>
<td><strong>LEP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-2.12</td>
<td>-2.05</td>
<td>-2.09</td>
<td>0.22</td>
<td>2.24</td>
<td>-1.14</td>
<td>15.62</td>
<td>6.85</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.24</td>
<td>0.52</td>
<td>0.33</td>
<td>2.39</td>
<td>3.70</td>
<td>1.08</td>
<td>3.08</td>
<td>3.07</td>
</tr>
<tr>
<td><strong>FRL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-4.68</td>
<td>-3.61</td>
<td>-3.95</td>
<td>3.48</td>
<td>5.68</td>
<td>-3.74</td>
<td>18.55</td>
<td>11.27</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.61</td>
<td>2.32</td>
<td>1.96</td>
<td>5.38</td>
<td>5.61</td>
<td>1.74</td>
<td>2.78</td>
<td>5.44</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 1. American Indian within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

The American Indian cluster profile analysis can be found in figures 1 and 2. As figure 1 illustrates, the American Indian subgroup has an average standardized value of 0.04 with clusters 2, 3, 5, 6, and 8 demonstrating mean values greater than average. Due to error bar overlap and small cluster averages, no cluster appears to demonstrate significant positive mean values. Figure 2 supports this hypothesis, with clusters 2, 3, 5, 6, and 8 not displayed due to a lack of significant in the positive direction at $\alpha = .05$. An evaluation of the American Indian within-cluster variation chart, in conjunction with the subgroups clusterwise importance plot, would indicate cluster 5 represents those schools.
with a large average, but statistically insignificant, population of American Indian students.

Figure 2. American Indian Clusterwise Importance

![American Indian Z-Score](image)

The cluster variation and clusterwise importance plots for the Asian subgroup is found in figures 3 and 4, respectively. As figure 3 illustrates, the Asian subgroup generated an overall average value of 0.40 with cluster 6 positive and significant (n = 12, $M = 5.35, SD = 2.79$). A similar result is found in the clusterwise importance chart (figure 4), with cluster 6 positive and significant ($t(11) \approx 7, p < .05$). Evaluating the
within-cluster variation chart, along with clusterwise importance plot, would indicate the 12 schools making up cluster 6 represent a group of schools having a significantly larger population of Asian students.

Figure 3. Asian within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = .4038

The Hispanic subgroup, with an average value of 2.02, appears to have two unique and significant clusters. As figure 5 illustrates, cluster 7 (n = 17, M = 16.03, SD = 2.88) and cluster 8 (n = 30, M = 7.92, SD = 3.26) appear positive and significant. The clusterwise importance plot (figure 6) confirms the existence of two unique and
significant clusters, with cluster seven $t(16) \approx 21, p < .05$ and cluster 8 $t(29) \approx 12, p < .05$).

Figure 4. Asian Clusterwise Importance

![Asian Z-Score](image)

Figures 7 and 8 represent the within cluster variation chart and clusterwise importance plot, respectively, for the African American population. As figure 7 illustrates, the African American subgroup has an overall mean of 0.77 with cluster 4 ($n = 26, M = 5.24, SD = 7.52$) representing what appears a single significant cluster. Figure 8, the African American clusterwise importance plot, reinforces this analysis with cluster 4
identified as significant at the 95% confidence level ($t(25) \approx 3, p < .05$). Evaluating the African American within-cluster variation, in conjunction with the subgroups clusterwise importance plot, would indicate cluster 4 represents a unique group of 26 schools enrolling larger than average populations of African American students.

Figure 5. Hispanic within Cluster Variation

![Simultaneous 95% Confidence Intervals for Means](image)

The within cluster variation chart and clusterwise importance plot for the white subgroup identifies three similar and significant cluster groups. The cluster variation chart, figure 9, quantifies a mean value of 1.67 and identifies cluster 1 ($n = 19, M = $
10.51, $SD = 1.94$), cluster 2 ($n = 25, M = 2.73, SD = 2.39$) and cluster 3 ($n = 18, M = 8.70, SD = 2.99$) as unique and significant. This is further verified with the clusterwise importance chart, figure 10, identifying cluster 1 ($t(18) = 20, p < .05$), cluster 2 ($t(24) = 16, p < .05$), and cluster 3 ($t(17) = 11, p < .05$) as statistically significant at the .05 level of significance.

Figure 6. Hispanic Clusterwise Importance

Hispanic Z-Score

Bonferroni Adjustment Applied

Critical Value

Test Statistic

Non-significant
Figure 7. African American within Cluster Variation

Moving into special populations, figure 11 characterizes the within cluster variation plot for students with individualized education plans (IEP). As figure 11 illustrates, students with IEP's have an overall mean value of 0.08 with cluster 3 (n = 18, $M = 0.53$, $SD = 0.40$) and cluster 5 (n = 12, $M = 1.44$, $SD = 0.60$) unique and significant. This is further verified with the IEP clusterwise importance chart, figure 12, identifying cluster 3 ($t(17) \approx 4$, $p < .05$) and cluster 5 ($t(11) \approx 7$, $p < .05$) as unique and statistically significant.
Students with limited English proficiency (LEP) had an overall mean value of 2.27 and identified cluster 7 (n = 17, $M = 15.62$, $SD = 3.08$) and cluster 8 (n = 30, $M = 6.85$, $SD = 3.07$) as unique and significant (figure 13). Figure 14, the LEP clusterwise importance chart, corroborates the significance of cluster 7 ($t(16) \approx 17, p < .05$) and cluster 8 ($t(29) \approx 9, p < .05$). This follows a pattern similar to the Hispanic subgroup, with the Hispanic average 2.02 and significant with clusters 7 (n = 17, $M = 16.06$, $SD = 2.88$, $t(16) \approx 21, p < .05$) and 8 (n = 30, $M = 7.92$, $SD = 3.26$, $t(29) \approx 12, p < .05$).

Should it be determined a strong multicollinearity exists between the Hispanic and LEP
subgroups the LEP subgroup may be represented by the Hispanic subgroup in future analysis.

Figure 9. White within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 1.6706
Figure 10. White Clusterwise Importance

White Z-Score

Bonferroni Adjustment Applied

- Critical Value
- Test Statistic

Cluster

Student's t

(0) Non-significant

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 11. IEP within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 0.047

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 12. IEP Clusterwise Importance

IEP Z-Score

Bonferroni Adjustment Applied

Critical Value

Test Statistic

(Student's t) Non-significant

72
Figure 13. LEP within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 2.0000
Figure 14. LEP Clusterwise Importance

LEP Z-Score

Bonferroni Adjustment Applied

Critical Value

Test Statistic

Cluster

Student's t

(II) Non-significant

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 15. FRL within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 3.2530
And lastly, students eligible for free or reduced priced lunches (FRL), commonly used to gauge level of poverty, have an overall mean value of 3.24 and represent significant averages in cluster 7 (n = 17, M = 18.55, SD = 2.78) and cluster 8 (n = 30, M = 11.28, SD = 5.44) (figure 15). This significance is further confirmed with the clusterwise importance plot, figure 16, demonstrating as significant clusters 7 (t(16) ≈ 21, p < .05) and 8 (t(29) ≈ 9, p < .05). It appears students classified FRL, much like students classified limited English proficient, demonstrate classification characteristics similar to
the Hispanic subgroup. As with LEP, should strong multicollinearity exists between Hispanic and FRL the FRL subgroup will be represented by the Hispanic subgroup.

First Cluster Analysis Summary

The American Indian subgroup, representing approximately 0.80% of the assessed population, was included in the initial phase to eliminate questions that may arise concerning subgroup significance. In practice, when inference testing proportions the population must be at least 10 times larger than the sample and $n_1 \hat{p}_1, n_1 (1 - \hat{p}_1), n_2 \hat{p}_2$, and $n_2 (1 - \hat{p}_2)$ greater than or equal to 5 (Hinkle, Wiersma, & Jurs, 1998). With no elementary schools included in this study having a large enough American Indian population to satisfy the minimum requirements for inclusion in inference testing, the American Indian subgroup was eliminated from further analysis.

Also of concern is the appearance of strong multicollinearity between Hispanic, LEP and FRL subgroups (Figures 5, 6, 13-16). As table 4 illustrates, the Pearson’s product-moment correlation coefficient between the seven identified subgroups (with American Indian eliminated from analysis) identifies a strong positive linear relationship between Hispanic, LEP, and FRL. The Hispanic to LEP correlation coefficient of .98 ($r = .98$) results in a .96 coefficient of determination ($r^2 = .96$), denoting 96% of the variation in LEP can be explained by variations in the Hispanic population (or vice versa). An equally strong relationship of $r = .86$ between Hispanic and FRL signifies 74% of the variation in FRL can be explained by variations in the Hispanic populations. And finally, the correlation coefficient between LEP and FRL, $r = .84$, equates to 71% of the variation in one subgroup being explainable by variations in the other.
Table 4. Pearson’s Correlation Coefficient of Subgroup Z-Score

<table>
<thead>
<tr>
<th></th>
<th>Asian</th>
<th>Hispanic</th>
<th>Af. Am.</th>
<th>White</th>
<th>IEP</th>
<th>LEP</th>
<th>FRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td>1</td>
<td>-.36</td>
<td>-.10</td>
<td>.07</td>
<td>-.25</td>
<td>-.29</td>
<td>-.39</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-.36</td>
<td>1</td>
<td>-.16</td>
<td>-.67</td>
<td>-.13</td>
<td>.98</td>
<td>.86</td>
</tr>
<tr>
<td>Af. Am.</td>
<td>-.10</td>
<td>-.16</td>
<td>1</td>
<td>-.24</td>
<td>.12</td>
<td>-.18</td>
<td>.16</td>
</tr>
<tr>
<td>White</td>
<td>.07</td>
<td>-.67</td>
<td>-.24</td>
<td>1</td>
<td>.02</td>
<td>-.61</td>
<td>-.71</td>
</tr>
<tr>
<td>IEP</td>
<td>-.25</td>
<td>-.13</td>
<td>.12</td>
<td>.02</td>
<td>1</td>
<td>-.19</td>
<td>-.05</td>
</tr>
<tr>
<td>LEP</td>
<td>-.29</td>
<td>.98</td>
<td>-.18</td>
<td>-.61</td>
<td>-.19</td>
<td>1</td>
<td>.84</td>
</tr>
<tr>
<td>FRL</td>
<td>-.39</td>
<td>.86</td>
<td>.16</td>
<td>-.71</td>
<td>-.05</td>
<td>.84</td>
<td>1</td>
</tr>
</tbody>
</table>

With this strong positive correlation between Hispanic, LEP, and FRL, it may be appropriate to designate any of the three as representative of the others. For example, it may be appropriate to assume all measurements and variations related to the LEP and FRL populations can be represented by the Hispanic subgroup. However, as a cautionary measure, it would be appropriate to first confirm LEP and FRL variations are related to Hispanic only. To substantiate this relationship, four additional TwoStep cluster analysis procedures were performed, each generating three unique clusters to represent three discrete variables, examining the relationships between Asian, LEP, & FRL; Hispanic, LEP, and FRL; African American, LEP, and FRL; and White, LEP, and FRL.

As figure 17 illustrates, the Asian subgroup has an average value of 0.40 with cluster 1 positive and significant. This is further confirmed by examining figure 18, the
Asian clusterwise importance chart, with cluster 1 positive and significant ($t(45)\approx 5, p < .05$). The LEP within-cluster variation chart, figure 19, as well as the FRL within-cluster variation chart, figure 21, demonstrates a positive and significant relationship with cluster 3. This relationship is further verified with the LEP and FRL clusterwise importance charts, figures 20 and 22, demonstrating a positive, significant relationship with cluster 3 that includes LEP ($t(54) \approx 8, p < .05$) and FRL ($t(54) \approx 20, p < .05$). The cluster distribution chart for Asian, LEP, and FRL is found in table 5 with cluster centroids in table 6. As this simple three cluster, TwoStep procedure verified, LEP and FRL were identified as similar and significant with Asian forming a unique significant cluster, significant within the Asian subgroup only and not statistically tied to the LEP or FRL subgroups.
Figure 17. Asian, with LEP and FRL within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = .4033
Figure 18. Asian, with LEP and FRL Clusterwise Importance

Asian Z Score

Bonferroni Adjustment Applied

Critical Value

Test Statistic

Cluster

Student's t

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 19. LEP, with Asian and FRL within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 2.2680
Figure 20. LEP, with Asian and FRL Clusterwise Importance

LEP Z-Score

Bonferroni Adjustment Applied

--- Critical Value

Test Statistic

Cluster

Student's t

83
Figure 21. FRL, with Asian and LEP within Cluster Variation

Simultaneous 95\% Confidence Intervals for Means

Reference Line is the Overall Mean = 3.5530
Table 5. Asian, LEP, and FRL Cluster Distribution

<table>
<thead>
<tr>
<th>Cluster</th>
<th>N</th>
<th>% of Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46</td>
<td>28.93%</td>
</tr>
<tr>
<td>2</td>
<td>58</td>
<td>36.48%</td>
</tr>
<tr>
<td>3</td>
<td>55</td>
<td>34.59%</td>
</tr>
<tr>
<td>Total</td>
<td>159</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Figure 22. FRL, with Asian and LEP Clusterwise Importance
Table 6. Asian, LEP, and FRL Cluster Centroids

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASIAN</td>
<td>Mean</td>
<td>2.23</td>
<td>-0.27</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>2.42</td>
<td>0.28</td>
</tr>
<tr>
<td>LEP</td>
<td>Mean</td>
<td>-1.44</td>
<td>-0.98</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>1.84</td>
<td>1.79</td>
</tr>
<tr>
<td>FRL</td>
<td>Mean</td>
<td>-3.46</td>
<td>-1.59</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>1.97</td>
<td>3.67</td>
</tr>
</tbody>
</table>

Hispanic, LEP, and FRL

The Hispanic, LEP, and FRL within-cluster variations and clusterwise importance charts for the TwoStep procedure that includes the subgroups Hispanic, LEP, and FRL can be found in figure 23–28 with cluster distributions and cluster centroids in tables 7 and 8, respectively. As expected, Hispanic, LEP, and FRL have been identified as a single cluster; cluster 1, representing 46 schools with overall centroid means of 11.46 for Hispanic, 10.53 for LEP, and 14.46 for FRL. This relationship between Hispanic, LEP, and FRL is further confirmed by the studentized t-values between cluster 1 and Hispanic (figure 24, t(45) ≈ 17, p < .05), LEP (figure 26, t(45) ≈ 12, p < .05), and FRL (figure 28, t(45) ≈ 14, p < .05). While this relationship was first observed and verified in the original TwoStep procedure, examining the TwoStep characteristics with only Hispanic, LEP and FRL further establishes the strong relationship between this unique collections of variables.
Figure 23. Hispanic, with LEP and FRL within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 2.0302
Figure 24. Hispanic, with LEP and FRL Clusterwise Importance

Hispanic Z-Score

Bonferroni Adjustment Applied

Cluster

Student's t

(H) Non-significant
Figure 25. LEP, with Hispanic and FRL within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 3.2680
Figure 26. LEP, with Hispanic and FRL Clusterwise Importance

LEP Z-Score

Bonferroni Adjustment Applied

--- Critical Value

Test Statistic

Cluster

Student's t

-60 -40 -20 0 20 30

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 27. FRL, with Hispanic and LEP within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 3.2530
Figure 28. FRL, with Hispanic and LEP Clusterwise Importance

FRL Z-Score

Bonferroni Adjustment Applied

Critical Value

Test Statistic

Student's t

Cluster

92

Non-significant

Table 7. Hispanic, LEP, and FRL Cluster Distributions

<table>
<thead>
<tr>
<th>Cluster</th>
<th>N</th>
<th>% of Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46</td>
<td>28.93%</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>20.75%</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>50.31%</td>
</tr>
<tr>
<td>Total</td>
<td>159</td>
<td>100.00%</td>
</tr>
</tbody>
</table>
Table 8. Hispanic, LEP, and FRL Cluster Centroids

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HISPANIC</td>
<td>11.46</td>
<td>1.29</td>
<td>-3.11</td>
</tr>
<tr>
<td>LEP</td>
<td>10.53</td>
<td>0.93</td>
<td>-1.93</td>
</tr>
<tr>
<td>FRL</td>
<td>14.46</td>
<td>5.12</td>
<td>-3.96</td>
</tr>
<tr>
<td><strong>Std. Deviation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HISPANIC</td>
<td>4.48</td>
<td>2.43</td>
<td>0.74</td>
</tr>
<tr>
<td>LEP</td>
<td>4.80</td>
<td>2.15</td>
<td>0.66</td>
</tr>
<tr>
<td>FRL</td>
<td>5.38</td>
<td>4.47</td>
<td>1.42</td>
</tr>
</tbody>
</table>

African American, LEP, and FRL

As the African American within cluster variation chart (figure 29) and clusterwise importance chart (figure 30) illustrate, when performing the TwoStep Cluster procedure using variables African American, LEP and FRL, the African American subgroup is identified in cluster 2 as unique and significant ($t(37) \approx 7, p < .05$). LEP and FRL are identified as similar (cluster one) with an LEP mean value of 9.00 and studentized $t(56) \approx 9, p < .05$ and an FRL mean value of 12.63 with a studentized $t(56) \approx 11, p < .05$. Table 9 illustrates the African American, LEP and FRL cluster distribution with table 10 listing the cluster centroids. As figure 29-34 and tables 9-10 confirm, the African American subgroup is unique and independent from LEP and FRL with LEP and FRL strongly correlated through a single, statistically significant group.
Figure 29. African American, with LEP and FRL within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = .7717
Figure 30. African American, with LEP and FRL Clusterwise Importance

African American Z-Score

Bonferroni Adjustment Applied

Critical Value

Test Statistic

Cluster

Student's t

-30 -20 -10 0
Figure 31. LEP, with African American and FRL within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 2.26

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 32. LEP, with African American and FRL Clusterwise Importance

LEP Z-Score

Bonferroni Adjustment Applied

- Critical Value
□ Test Statistic

Cluster

-30 -20 -10 0 10

Student's t

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 33. FRL, with African American and LEP within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 3.2530
Figure 34. FRL, with African American and LEP Clusterwise Importance

FRL Z-Score

Table 9. African American, LEP, and FRL Cluster Distributions

<table>
<thead>
<tr>
<th>Cluster</th>
<th>N</th>
<th>% of Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>57</td>
<td>35.85%</td>
</tr>
<tr>
<td>2</td>
<td>38</td>
<td>23.90%</td>
</tr>
<tr>
<td>3</td>
<td>64</td>
<td>40.25%</td>
</tr>
<tr>
<td>Total</td>
<td>159</td>
<td>100.00%</td>
</tr>
</tbody>
</table>
Table 10. *African American, LEP, and FRL Cluster Centroids*

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>African American</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-0.16</td>
<td>4.47</td>
<td>-0.59</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>1.05</td>
<td>6.29</td>
<td>0.43</td>
</tr>
<tr>
<td><strong>LEP</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>9.00</td>
<td>-1.04</td>
<td>-1.77</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>5.38</td>
<td>1.52</td>
<td>1.11</td>
</tr>
<tr>
<td><strong>FRL</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>12.63</td>
<td>1.55</td>
<td>-4.08</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>6.39</td>
<td>5.55</td>
<td>1.65</td>
</tr>
</tbody>
</table>

Figure 35. White, with LEP and FRL within Cluster Variation

*Simultaneous 95% Confidence Intervals for Means*

Reference Line is the Overall Mean = 1.6706
White, LEP, and FRL

The TwoStep cluster analysis of White, LEP, and FRL is the final measure in the verification of a statistical relationship between Hispanic, LEP, and FRL. As figures 35 and 36 illustrate, cluster 1 is unique and statistically significant in identifying the white subgroup with a mean of 8.65 and a studentized \( t(48) \approx 17, p < .05 \). LEP and FRL have again been identified as a single group (cluster two) with an LEP mean of 10.82 and studentized \( t(65) \approx 12, p < .05 \) and an FRL mean of 14.69 with studentized \( t(65) \approx 17, p < .05 \). The White, LEP, and FRL cluster distribution chart can be found in table 11 with cluster centroids chart in table 12.

Figure 36. White, with LEP and FRL Clusterwise Importance

White Z-Score

![White Z-Score Diagram](image-url)
Figure 37. LEP, with White and FRL within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 2.2680
Figure 38. LEP, with White and FRL Clusterwise Importance
Figure 39. FRL, with White and LEP within Cluster Variation

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 3.2530
Figure 40. FRL, with White and LEP Clusterwise Importance

Table 11. White, LEP, and FRL Cluster Distributions

<table>
<thead>
<tr>
<th>Cluster</th>
<th>N</th>
<th>% of Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>49</td>
<td>30.82%</td>
</tr>
<tr>
<td>2</td>
<td>66</td>
<td>41.51%</td>
</tr>
<tr>
<td>3</td>
<td>44</td>
<td>27.67%</td>
</tr>
<tr>
<td>Total</td>
<td>159</td>
<td>100.00%</td>
</tr>
</tbody>
</table>
Evidence of strong multicollinearity between the Hispanic, LEP and FRL subgroups first appeared during the initial TwoStep cluster analysis phase and was further confirmed by Pearson’s product-moment correlation coefficient (table 4). As a cautionary measure, the LEP and FRL subgroups were analyzed separately, using the same TwoStep cluster analysis procedure, against the Asian, Hispanic, African American, and White subgroups. Final results confirm the Asian, African American, and White subgroups are unique and independent from LEP and FRL with Hispanic, LEP and FRL exhibiting strong multicollinearity. The presence of multicollinearity results in redundant information, making it unnecessary to include all three subgroups in the final analysis. Therefore, for the remainder of this study, the LEP and FRL subgroups will be dropped from the data set and the Hispanic subgroup used to represent the Hispanic, LEP, and FRL subgroups.

Phase II: Asian, Hispanic, African American, White, and LEP

With the American Indian subgroup removed from analysis due to insignificant populations and Hispanic representative of LEP and FRL, a second phase of TwoStep
cluster analysis was performed to analyze the relationships between the Asian, Hispanic, African American, White, and IEP subgroups. A summary of cluster distributions, found in table 13, reveals a membership low of 14 schools in cluster 4 to 52 schools in cluster 5, with cluster centroid means and standard deviations found in table 14. Actual cluster membership, by school, can be found in Appendix E.

Analyzing table 14, along with Appendix E, reveals patterns that provide preliminary information with respect to cluster distributions. Asian quantifies a maximum centroid cluster mean of 3.19 in cluster 3, with the 28 schools in cluster 3 having a combined enrollment mean of 3.19 standard deviations above district average of enrolled Asian students. This pattern follows with the Hispanic subgroup generating a maximum mean value of 10.02 for the 52 schools in cluster 5, African American a mean of 10.70 for the 14 cluster 4 schools, White with a mean value of 8.71 for cluster 2, and IEP an overall mean value of 0.88 standard deviations above district average for cluster 1. This preliminary observation of cluster membership may prove accurate should the respective cluster distributions withstand significance testing.

The Asian within cluster variation chart, figure 41, identifies cluster 3 as having a positive, significant mean value of 3.19 with a standard deviation of 2.68 (table 14). This significant relationship is further verified in the Asian clusterwise importance chart, found in figure 42, which identifies cluster 3 as a single, statistically significant cluster for the Asian subgroup with a studentized value of $t(27) \approx 8, p < .05$. 

107
### Table 13. Phase II Cluster Distributions

<table>
<thead>
<tr>
<th>Cluster</th>
<th>N</th>
<th>% of Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>22.01%</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>18.87%</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>17.61%</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>8.81%</td>
</tr>
<tr>
<td>5</td>
<td>52</td>
<td>32.70%</td>
</tr>
<tr>
<td>Total</td>
<td>159</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

### Table 14. Phase II Cluster Centroids

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASIAN</td>
<td>Mean</td>
<td>-0.18</td>
<td>0.14</td>
<td>3.19</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>0.4</td>
<td>0.6</td>
<td>2.68</td>
<td>0.48</td>
</tr>
<tr>
<td>HISPANIC</td>
<td>Mean</td>
<td>-0.55</td>
<td>-3.36</td>
<td>-1.88</td>
<td>-1.92</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>4.02</td>
<td>0.63</td>
<td>2.72</td>
<td>0.91</td>
</tr>
<tr>
<td>AFRICAN</td>
<td>Mean</td>
<td>0.02</td>
<td>-0.59</td>
<td>-0.31</td>
<td>10.7</td>
</tr>
<tr>
<td>AMERICAN</td>
<td>Std. Dev.</td>
<td>1.04</td>
<td>0.54</td>
<td>0.54</td>
<td>6.64</td>
</tr>
<tr>
<td>WHITE</td>
<td>Mean</td>
<td>3.79</td>
<td>8.71</td>
<td>1.54</td>
<td>-1.68</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>5.21</td>
<td>3.13</td>
<td>3.02</td>
<td>0.79</td>
</tr>
<tr>
<td>IEP</td>
<td>Mean</td>
<td>0.88</td>
<td>-0.18</td>
<td>-0.26</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>0.58</td>
<td>0.29</td>
<td>0.32</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Figure 41. Asian within Cluster Variation, Phase II

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = -4.038
Figure 42. Asian Clusterwise Importance, Phase II

Asian Z-Score

Bonferroni Adjustment Applied

Critical Value

Test Statistic

Cluster

Student's t

(M) Non-significant
Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 2.0302
The Hispanic cluster variation chart, which is representative of Hispanic, LEP, and FRL, can be found in figure 43. Cluster 5 appears to represent this group with a positive centroid mean value of 10.02 and standard deviation of 5.44 (table 14). Figure 44, the Hispanic clusterwise importance chart, supports this conclusion, identifying cluster 5 as the single, statistically significant cluster for Hispanic, LEP, and FRL with a studentized $t(51) = 12, p < .05$.

Cluster 4 appears to be statistically significant with respect to variable African American, having a centroid mean value of 10.70 and standard deviation 6.64 (table 14,
The clusterwise importance chart, figure 46, also identifies cluster 4 as unique and significant with a studentized $t(13) \approx 7, p < .05$. This would suggest the 14 schools included in cluster 4 enroll African American populations that are, on average, $10.70$ standard deviations above district average.

Figure 47, the White within cluster variation chart, identifies clusters 1 and 2 as representative of the White subgroup, having a cluster 1 mean value of 3.79 with a standard deviation of 5.21 and a cluster 2 mean of 8.71 with a 3.13 standard deviation (table 14, figure 47). While two positive centroid mean values are identified for the White subgroup, the White clusterwise importance chart, figure 48, identifies cluster 2 as the single, statistically significant cluster with a studentized $t(29) \approx 15, p < .05$.

And lastly, the IEP subgroup cluster variation chart, figure 49, identifies two positive mean centroid values with cluster 1 generating a mean centroid value of 0.88 with a 0.58 standard deviation and cluster 4 having a mean centroid value of 0.14 with a 0.26 standard deviation (table 14). Although two positive centroid mean values exist, the clusterwise importance chart, figure 50, identifies cluster 1 as the single, statistically significant cluster representative of those schools enrolling larger than average percentages of IEP students with a studentized $t(34) \approx 8, p < .05$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 45. African American within Cluster Variation, Phase II

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = .7717
Figure 46. African American Clusterwise Importance, Phase II

African American Z-Score

Bonferroni Adjustment Applied

- Critical Value
- Test Statistic

Cluster

Student's t

115

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 47. White within Cluster Variation, Phase II

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = 1.6706

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 48. White Clusterwise Importance, Phase II

**White Z-Score**

Clusterwise Importance, Phase II

**Bonferroni Adjustment Applied**

Critical Value

Test Statistic

(Student's t) Non-significant
Figure 49. IEP within Cluster Variation, Phase II

Simultaneous 95% Confidence Intervals for Means

Reference Line is the Overall Mean = .0847
In summary, the phase II TwoStep cluster analysis procedure defined and generated five unique and statistically significant clusters of schools using subgroup variables Asian, Hispanic, African American, White, and IEP. This would suggest that while NCLB requires Clark County School District to report the performance for eight separate subgroups that include American Indian, Asian, Hispanic, African American, White, IEP, LEP, and FRL, the subgroups significant enough to provide valid results would include Asian, African American, White, IEP, and a combined Hispanic, LEP and FRL subgroup.
Questions 2 and 3: Does a statistically significant difference in mean reading or mean math CRT scale scores exist within homogeneous clusters of five or more schools?

NCLB requires schools report the performance for any subgroup containing 25 or more students. As the TwoStep cluster analysis procedure has demonstrated, schools identified as having homogeneous populations from any of the eight identified subgroups simply reports varying group sizes within that single homogeneous group. For example, Tom Williams Elementary School had 384 students fulfilling the requirements for inclusion in this study. And of that 384, zero were American Indian, two Asian, 359 Hispanic, five African American, 18 white, 50 IEP, 301 LEP, and 383 FRL. All subgroup reports from Tom Williams ES would simply evaluate a sample from those core 356 Hispanic students. The 384 students included in the school wide analysis would, statistically speaking, represent the same group evaluated in the Hispanic subgroup. The American Indian, Asian, African American, and White populations were statistically insignificant due to small group size and would not report. The remaining 50 IEP, 301 LEP, and 383 FRL were again sample groups from the 356 Hispanic students. Although Tom Williams ES would be evaluated in five separate areas for ELA and five separate areas for math, increasing the probability of a type II error with each additional evaluation, the final measurement simply quantifies the ELA and math performance for the 356 Hispanic students.
African American Subgroup Analysis

The African American subgroup formed a unique, homogeneous cluster of 14 schools contained in cluster 4, which from this point forward will be referred to as the African American cluster. The list of schools and z-scores for the African American cluster can be found in Appendix E. Schools in this cluster were Booker ES, Carson ES, Elizondo ES, Gilbert ES, Hoggard ES, Mackey ES, Priest ES, Reed ES Watson ES, Wilhelm ES, and Wendell Williams ES.

Before analysis could begin, it was necessary to identify and eliminate all outliers that might exist in the mean reading or mean math scale scores for the African American cluster. A school identified as an outlier would have a mean reading or mean math scale score abnormally larger or smaller than expected, thereby influencing the distribution of mean scale scores. Using the definition of outlier as a mean scale score outside the 1.5 x inner quartile range (IQR), the reading quartile 1 (Q1) of 245.63, median of 259.13, quartile 3 (Q3) of 275.12, (IQR) of 29.49 (Q1 – Q3), and 1.5 x IQR of 44.24 resulted in extreme values of 201.39 and 319.36. With the minimum mean reading scale score of 217.36 (Kelly ES) and maximum mean reading scale score of 302.76 (Hoggard ES), no African American schools were identified as having outlier mean reading scale scores (Appendix F).

Repeating the outlier calculations with mean math scale scores for the African American cluster resulted in a Q1 of 257.31, median of 262.88, Q3 of 272.56, IQR of 15.25, and 1.5 x IQR of 22.89. The extreme mean math scale scores of 234.42 and 295.45 identified Kelly ES (M = 221.28) and Fitzgerald ES (M = 229.60) as outliers with mean math scale scores below expected distributions and Hoggard ES (M = 311.95) as an outlier.
outlier with a mean math scale score above expected distributions. Based upon this
group of calculations, for analysis purposes all schools were included in the reading
ANOVA study while Kelly ES and Fitzgerald ES were removed from math analysis due
to math scale scores significantly below expected distributions and Hoggard ES removed
from math analysis with mean math scale scores significantly above expected
distributions.

The one-way Analysis of Variance (ANOVA), using an f-test of difference in
mean reading and mean math scale scores against categorical variable school name,
determined if the mean reading or mean math scale score for individual schools within
the African American cluster differed significantly with respect to comparison African
American cluster school mean reading and mean math scale scores. Methods for analysis
was dependent upon Levene’s test of homogeneity of variances, a procedure designed to
test the null hypothesis all group variances are equal. If the Levene output p-value is
greater than .05, equal variances can be assumed and standard ANOVA procedures
carried out using Tukey’s HSD post-hoc analysis. In such cases where the assumption of
homogeneity of variances is violated (p < .05), the Welch robust tests of equality of
means $F$-ratio will be reported and post-hoc analysis carried out using Tamhane’s T2
procedure.

The main ANOVA summary table and Levene’s test of homogeneity of variances
for the African American cluster mean reading and mean math scale scores can be found
in tables 15 and 16, respectively. With Levene’s statistic less than .05 for reading and
math, equality of variances could not be assumed, making it necessary to report the $F$-
ratio of the alternative statistic. Welch’s $F$-ratio, found in table 17, revealed the mean
reading and mean math scale scores differ significantly as a function of school with reading \( F(13,551.77) = 11.06, p < .05 \), and math \( F(10,474.36) = 2.90, p < .05 \). For reference purposes, school level reading and math mean scale scores, standard deviations, standard errors, 95% confidence intervals, and minimum/maximum values can be found in Appendix F.

To determine which schools experienced significant differences in mean reading and mean math scale scores the post-hoc Tamhane’s T2 procedure for unequal variances was used. A conservative pairwise comparison test based upon the t-procedure and commonly used when encountering unequal variances, Tamhane’s T2 post-hoc procedure compared the differences in mean reading and mean math scale score between school (I) and each comparison school (J) within the African American cluster to generate a matrix of differences (I – J), using an asterisk to indicate significance at \( \alpha = .05 \) (Appendix G). Also included in Tamhane’s T2 matrix are standard error of the differences, level of significance (\( \alpha = .05 \)), and 95% confidence interval of the differences.

<table>
<thead>
<tr>
<th>Table 15. African American ANOVA Summary Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of Mean Squares</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>RSS</td>
</tr>
<tr>
<td>Between Groups</td>
</tr>
<tr>
<td>Within Groups</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td>MSS</td>
</tr>
<tr>
<td>Between Groups</td>
</tr>
<tr>
<td>Within Groups</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

123
Table 16. **Levene’s Test of Homogeneity of Variance for African American**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS</td>
<td>2.20</td>
<td>13</td>
<td>1553</td>
</tr>
<tr>
<td>MSS</td>
<td>2.53</td>
<td>10</td>
<td>1264</td>
</tr>
</tbody>
</table>

Table 17. **Welch Robust Tests of Equality of Means for African American**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS</td>
<td>11.06</td>
<td>13</td>
<td>551.77</td>
</tr>
<tr>
<td>MSS</td>
<td>2.90</td>
<td>10</td>
<td>474.36</td>
</tr>
</tbody>
</table>

The differences in mean reading scale scores for those schools located in the African American cluster identified Hoggard ES as having the greatest mean reading scale score, \( M = 302.76 \) (Appendix F), that was significantly greater than the mean reading scale score for Wendell Williams ES (\( M = 252.34 \)), Priest ES (\( M = 252.07 \)), Tartan ES (\( M = 245.63 \)), Elizondo ES (\( M = 238.49 \)), Fitzgerald ES (\( M = 220.62 \)), and Kelly ES (\( M = 217.36 \)). Booker ES (\( M = 284.03 \)) and Mackey ES (\( M = 279.01 \)) had the second and third highest mean reading scale scores and were also significantly greater than Tartan, Elizondo, Fitzgerald, and Kelly Elementary Schools, but, statistically speaking, not unlike the remaining schools identified in the African American cluster. This lack of significance in mean reading scale scores demonstrates the homogeneity of those schools identified as having significantly large African American populations.

Those schools demonstrating significantly lower mean reading scale scores included Wendell Williams ES, Priest ES and Watson ES (\( M = 251.78 \)), each significantly lower than Hoggard ES. Other African American schools identified as
having significantly lower mean reading scores were Tartan ES, with a mean reading scale score that was significantly less than Hoggard ES, Booker ES, and Mackey ES with Elizondo ES significantly less than Hoggard ES, Booker ES, Mackey ES and Gilbert ES. And finally, Fitzgerald ES and Kelly ES each had mean reading scale scores that were significantly less than all elementary schools located in the African American cluster with the exception of Tartan ES and Elizondo ES. To summarize, Hoggard ES, Booker ES, and Mackey ES appeared to have significantly greater mean reading scale scores with Tartan ES, Elizondo ES, Fitzgerald ES, and Kelly ES significantly lower. All others, which include Reed ES, Gilbert ES, Carson ES, Wilhelm ES, W. Williams ES, Priest ES, and Watson ES, appeared to have mean scale scores not significantly different than expected.

An analysis of the African American mean math scale scores began with removal from analysis outlier schools Hoggard ES (M = 311.95), Kelly ES (M = 221.28), and Fitzgerald ES (M = 229.60). From the remaining African American schools Gilbert ES (M = 281.97) and Mackey ES (M = 276.88) were identified as significant, having mean math scale scores that were significantly greater than Tartan ES (M = 239.02). All other schools in the African American cluster, which included Carson ES (M = 272.56), Booker ES (M = 268.20), Priest ES (M = 264.71), Watson ES (M = 263.69), Wilhelm ES (M = 262.06), Reed ES (M = 260.42), Elizondo ES (M = 258.19), and W. Williams ES (M = 257.31) had mean math scale scores that were, statistically speaking, not unalike. This similarity in mean math scale scores again demonstrates the homogeneity of the African American Cluster. To summarize the distribution of mean math scale scores in the African American cluster, Hoggard ES was identified as an outlier with a mean math
scale score above expected distributions, Kelly ES and Fitzgerald ES identified as outliers with mean math scale scores below expected distributions, Gilbert ES and Mackey ES significantly greater than Tartan ES, and all remaining African American schools with similar mean math scale scores.

**Asian Subgroup Analysis**

The Asian subgroup was identified as a unique, homogeneous cluster of 28 schools contained in cluster 3, which from this point forward will be referred to as the Asian cluster. The list of schools and z-scores for the Asian cluster can be found in Appendix E. This cluster of schools included Alamo ES, Bass ES, Batterman ES, Beatty ES, Bendorf ES, Roger Bryan ES, Cartwright ES, Decker ES, Diskin ES, Marion Earl ES, Frias ES, Gehring ES, Givens ES, Goolsby ES, Goynes ES, Hayes ES, Hummel ES, Iverson ES, Jydstrup ES, Kim ES, Mendoza ES, Ries ES, Rogers ES, Tanaka ES, Thiriot ES, Treem ES, Whitney ES, and Wolff ES.

Analysis began with identifying and eliminating any and all outliers from the Asian mean reading or mean math scale scores. Repeating the definition of outlier as a mean scale score outside the $1.5 \times IQR$, the reading Q1 of 307.93, median of 316.73, Q3 of 327.53, IQR of 19.6 ($Q1 - Q3$), and $1.5 \times IQR$ of 29.4 resulted in extreme values of 278.53 and 356.93. The lower-bound critical value of 278.53 would place the mean reading scale score for Thiriot ES ($M = 265.04$) outside the lower outlier range. No schools were identified as outliers in the upper-bound reading range with Goolsby ES having the maximum mean reading scale score ($M = 353.50$), well within the upper-bound critical value of 356.93.
Replicating the outlier analysis with the Asian mean math scale scores resulted in a Q1 of 326.81, median of 335.59, Q3 of 349.27, IQR of 22.46 (Q1 – Q3), and 1.5 x IQR of 33.69. The critical values of 293.12 and 382.96 identified Thiroit ES as an outlier below expected distributions (M = 282.60) and Bendorf ES an outlier above expected distributions (M = 391.41). Therefore, for analysis purposes Thiroit ES will be removed from the reading ANOVA study with Thiriot ES and Bendorf ES removed from math ANOVA analysis.

The one-way Analysis of Variance (ANOVA) was again used to determine if the mean reading and mean math scale scores within the Asian cluster differed significantly. With Levene’s statistic for reading $F(26,1940) = 1.14$, $P > .05$ and math $F(25,1859) = 0.96$, $P > 0.05$ (table 19), reading and math equality of variances could be assumed, allowing for standard ANOVA procedures using Tukey’s HSD post-hoc analysis. The standard ANOVA chart, found in table 18, revealed a significant relationship for reading, $F(26,1940) = 3.34$, $P < .05$, and math, $F(25,1859) = 3.72$, $P < .05$, allowing for rejection of the null hypothesis that all mean reading and math scale scores are the same and acceptance of the alternative hypothesis that at least two of the schools in the Asian cluster have significantly different mean reading or mean math scale scores. The Asian cluster mean scale scores, standard deviations, standard errors, 95% confidence intervals, and minimum/maximum values can be found in Appendix F.

To determine which schools experienced significant differences in mean reading and mean math scale scores, Tukey’s HSD post-hoc procedure for equal variances was used. Similar to Tamhane’s T2 post-hoc procedure, Tukey’s HSD is a conservative pairwise comparison test, based upon a t-procedure, that compares the differences in
mean reading and mean math scale score between school (I) and each comparison school (J) within the Asian cluster to generate a matrix of differences (I - J), using an asterisk to indicate significance at $\alpha = .05$ (Appendix G). Also included in Tukey's HSD matrix is the standard error of differences, level of significance ($\alpha = .05$), and 95% confidence interval of the differences.

An examination of the differences in mean reading scale scores for those schools located in the Asian cluster, less Thiroit ES, identified Bendorf ES ($M = 353.40$) as having a mean reading scale score significantly greater than the mean reading scale scores of Alamo ES ($M = 316.55$), Batterman ES ($M = 313.37$), Decker ES ($M = 309.54$), Diskin ES ($M = 293.94$), Gehring ES ($M = 299.27$), Hayes ES ($M = 305.65$), Hummel ES ($M = 312.12$), Jydstrup ES ($M = 308.13$), Ries ES ($M = 302.83$), Tanaka ES ($M = 307.73$), and Treem ES ($M = 311.03$, Appendix G). Goolsby ES ($M = 453.50$) had a mean reading scale score that was also significantly greater than Diskin ES, Gehring ES, Hayes ES, Ries ES, and Tanaka ES, with Givens ES ($M = 344.97$) significantly greater than Diskin ES and Gehring ES.

Evaluating the Asian schools with significantly lower mean reading scale scores revealed Alamo ES, Batterman ES, Decker ES, Hummel ES, Jydstrup ES, and Treem ES as having mean scale scores significantly lower than the scale score for Bendorf ES with Tanaka ES, Hayes ES, and Ries ES significantly lower than Bendorf ES and Goolsby ES. And finally, Gehring ES and Diskin ES each had mean reading scale scores that were significantly lower than Bendorf ES, Givens ES, and Goolsby ES. The remaining schools in the Asian cluster (Bass ES, $M = 320.08$; Beatty ES, $M = 321.64$; Roger Bryan ES, $M = 321.92$; Cartwright ES, $M = 328.78$; Marion Earl ES, $M = 316.90$; Frias ES, $M = 318.03$).
= 331.37; Goynes ES, M = 329.96; Iverson ES, M = 326.10; Kim ES, M = 326.28; Mendoza ES, M = 311.22; Rogers ES, M = 334.30; Whitney ES, M = 305.54; Wolff ES, M = 320.89) all had mean reading scale scores that were, statistically speaking, equivalent.

Table 18. **Asian ANOVA Summary Table**

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS Between Groups</td>
<td>370173.48</td>
<td>26</td>
<td>14237.44</td>
<td>3.34</td>
<td>0.00</td>
</tr>
<tr>
<td>Within Groups</td>
<td>8281274.03</td>
<td>1940</td>
<td>4268.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>8651447.52</td>
<td>1966</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSS Between Groups</td>
<td>436988.93</td>
<td>25</td>
<td>17479.56</td>
<td>3.72</td>
<td>0.00</td>
</tr>
<tr>
<td>Within Groups</td>
<td>8738661.32</td>
<td>1859</td>
<td>4700.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9175650.25</td>
<td>1884</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 19. **Levene's Test of Homogeneity of Variance for Asian**

<table>
<thead>
<tr>
<th></th>
<th>Levene Statistic</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS</td>
<td>1.14</td>
<td>26</td>
<td>1940</td>
<td>0.29</td>
</tr>
<tr>
<td>MSS</td>
<td>0.96</td>
<td>25</td>
<td>1859</td>
<td>0.52</td>
</tr>
</tbody>
</table>

The ANOVA procedure for Asian mean math scale scores began with removing from analysis outlier schools Thiroit ES (M = 282.60) and Bendorf ES (M = 391.41). An ANOVA procedure of the remaining schools, using Tukey’s HSD post-hoc analysis, identified Givens ES as having a mean math scale score (M = 378.22) significantly greater than Alamo ES (M = 338.01), Bass ES (M = 335.65), Batterman ES (M = 129
334.67), Decker ES (M = 324.53), Diskin ES (M = 314.58), Hummel ES (M = 322.05),
Iverson ES (M = 334.65), Jydstrup ES (M = 303.82), Mendoza ES (M = 322.36), Ries ES
(M = 329.09), Tanaka ES (M = 335.09), and Whitney ES (M = 316.75). Rogers ES (M =
368.87) produced a mean math scale score that was significantly greater than Decker ES,
Diskin ES, Hummel ES, and Jydstrup ES with Goolsby ES (M = 374.33) significantly
greater than Diskin ES, Hummel ES, and Jydstrup ES. And finally, Kim ES (M =
357.93) and Frias ES (M = 353.02) each had mean math scale scores significantly greater
than Jydstrup ES.

An evaluation of the mean math scale scores for the Asian cluster schools scoring
in the lower range revealed Alamo ES, Bass ES, Batterman ES, Iverson ES, Mendoza ES,
Ries ES, Tanaka ES, and Whitney ES as having mean math scale scores that were
significantly lower than Givens ES with Decker ES and Hummel ES both significantly
lower than Givens ES and Rogers ES. And finally, Diskin ES had a mean math scale
score that was significantly lower than Givens ES, Goolsby ES, and Rogers ES with
Jydstrup ES significantly lower than Givens ES, Goolsby ES, Frias ES, Kim ES, and
Rogers ES. The remaining Asian cluster schools (Beatty ES, M = 348.44; Roger Bryan
ES, M = 339.15; Cartwright ES, M = 350.10; Marion Earl ES, M = 336.20; Gehring ES,
M = 332.00; Goynes ES, M = 335.53; Hayes ES, M = 335.21; Treem ES, M = 336.59;
Wolff ES, M = 346.68) demonstrated no significant difference in mean math scale scores.

To summarize, Thiroit ES was identified as an outlier in the Asian cluster and
removed from analysis for both reading and math with Bendorf ES identified as an outlier
and removed for math only. Bendorf ES, Goolsby ES, and Givens ES were recognized as
having significantly greater reading scale scores with Givens ES, Rogers ES, Goolsby

130
ES, Kim ES, and Frias ES all having significantly greater mean math scale scores. Alamo ES, Batterman ES, Decker ES, Hummel ES, Jydstrup ES, and Treem ES had mean reading scale scores that were significantly lower than Bendorf ES with Hayes ES, Ries ES, and Tanaka ES significantly lower than Bendorf ES and Goolsby ES while Gehring ES and Diskin ES each had mean reading scale scores that were significantly lower than Bendorf ES, Givens ES, and Goolsby ES. Alamo ES, Bass ES, Batterman ES, Iverson ES, Mendoza ES, Ries ES, Tanaka ES, and Whitney ES all had mean math scale scores that were significantly lower than Givens ES with Decker ES significantly lower than Givens ES and Rogers ES while Hummel ES and Diskin ES each had mean math scale scores that were significantly lower than Givens ES, Goolsby ES, and Rogers ES. And finally, Jydstrup ES had a mean math scale score that was significantly lower than Givens ES, Goolsby ES, Frias ES, Kim ES, and Rogers ES. All other schools in the Asian cluster had mean reading and mean math scale scores that were, statistically speaking, equivalent.

**Hispanic Subgroup Analysis**

Cluster 5 formed a unique, homogeneous subgroup of 52 schools representative of the Hispanic subgroup, with a list of schools and z-scores found in Appendix E. In response to the effects of multicollinearity, the Hispanic cluster was representative of Hispanic, LEP, and FRL, making possible the inclusion of pairwise mutually exclusive schools for Hispanic, LEP, or FRL. To compensate for this misclassification, Cunningham ES (34.42% Hispanic, 22.22% LEP, & 66.67% FRL), Fong ES (42.68% Hispanic, 28.66% LEP & 62.42% FRL), and Paradise ES (47.13% Hispanic, 35.67%

As with the African American and Asian clusters, it was necessary to identify and eliminate from analysis any Hispanic mean reading or mean math scale score outliers. Using the definition of an outlier as any mean scale score outside the 1.5 x IQR, the Hispanic reading Q1 of 250.00, median of 260.86, Q3 of 264.75, IQR of 14.75 (Q1 – Q3), and 1.5 x IQR of 22.13 resulted in extreme values of 227.87 and 286.83. While no schools in the Hispanic cluster were found to have a mean reading scale scores below the lower critical value, Cahlan ES (M = 300.15) was identified as an outlier with a mean reading scale score above the upper critical value.

Outlier calculations for the mean math scale scores quantified a Q1 of 270.23, median of 276.62, Q3 of 286.29, IQR of 16.06 (Q1 – Q3), and 1.5 x IQR of 24.09, resulting in critical values of 246.14 and 310.28. Petersen ES (M = 234.13) was identified as an outlier below expected distributions with Cahlan ES (M = 314.95) identified as an outlier above expected distributions. To avoid bias resulting from the influence of outliers on scale score distributions, Cahlan ES was removed from the
Hispanic reading ANOVA study while Petersen ES and Cahlan ES were removed from the Hispanic math ANOVA analysis.

To determine if the mean reading and mean math scale scores for schools within the Hispanic cluster differed significantly, the one-way Analysis of Variance (ANOVA) was again performed using either Tukey's HSD or Tamhane's T2 post-hoc analysis. Methods for post-hoc analysis were once more dependent upon Levene's test of homogeneity of variances, with output p-values greater than .05 allowing Tukey's HSD while output p-values less than .05 violating the assumption of homogeneity of variances. In such cases where p < .05, the Welch robust tests of equality of means F-ratio was reported and post-hoc analysis carried out using Tamhane's T2 procedure. The main ANOVA summary table and Levene's test of homogeneity of variances for the Hispanic mean reading and mean math scale scores can be found in tables 20 and 21, respectively. With Levene's statistic less than .05 for both reading, $F(47,11434)=1.64$, $p < .05$, and math, $F(46,11263)=2.37$, $p < .05$, one of the assumptions of the main ANOVA had been violated, making it necessary to report the F-ratio of the alternative statistic, Welch's robust tests of equality of means. The Welch F-ratio, found in table 22, verified the mean reading and mean math scale scores differed significantly as a function of school with reading $F(47,3630.09) = 7.65$, $p < .05$, and math $F(46,3565.80) = 8.75$, $p < .05$. Hispanic reading and math mean scale scores, standard deviations, standard errors, 95% confidence intervals, and minimum/maximum values can be found in Appendix F with the post-hoc table of differences and significance levels in Appendix G.
The detailed analysis of the ANOVA procedure using mean reading scale scores across the Hispanic cluster schools, less outlier Cahlan ES with Tamhane’s T2 post-hoc procedure, resulted in Harmon ES (M = 285.64) having a mean reading scale score significantly greater that Bell ES (M = 242.90), Culley ES (M= 239.26), Craig ES (M = 252.48), Dearing ES (M = 253.34), Detwiler ES (M = 250.64), Ira Earl ES (M = 248.62),
Herron ES (M = 242.29), Hewetson ES (M = 258.23), Hollingsworth ES (M = 246.95), Jeffers ES (M = 231.96), Lake ES (M = 258.51), Long ES (M = 246.39), Lunt ES (M = 243.86), Manch ES (M = 236.41), Petersen ES (M = 229.96), Ronnow ES (M = 260.92), Sunrise Acres ES (M = 251.85), Tate ES (M = 242.26), Tom Williams ES (M = 254.97), and Wynn ES (M = 249.35). The next schools with significantly large reading scale scores were Crestwood ES (M = 283.05), Twin Lakes ES (M = 282.15) and Bracken ES (M = 179.91), each having a mean reading scale score significance level similar to Harmon ES less the significantly larger relationship with Dearing ES, Hewetson ES, Lake ES, and Ronnow ES. Lincoln ES followed next, having a mean reading scale score (M = 276.03) similar to Bracken ES, Crestwood ES, and Twin Lakes ES less the significantly greater relationship with Craig ES, Detwiler ES, Sunrise Acres ES, and Tom Williams ES. And finally, Red Rock ES (M = 277.84) had a mean reading scale score resembling Lincoln ES minus the significantly greater relationship to Wynn ES.

Edwards ES (M = 270.44) had a mean reading scale score that was significantly greater than Bell ES, Culley ES, Herron ES, Jeffers ES, Lunt ES, Manch ES, Peterson ES, and Tate ES with Vegas Verdes ES having a mean reading scale score (M = 271.04) similar to Edwards ES less the significantly greater relationship to Lunt ES and Tate ES. Snyder ES (M = 267.19) and Squires ES (M = 267.01) were also similar to Edwards less the significantly greater relationship with Culley ES while Dailey ES (M = 264.13) was similar to Snyder ES and Squires ES minus the significantly greater relationship with Bell ES. Gragson ES (M = 261.44), Lynch ES (M = 262.36), Park ES (M = 265.22), and Woolley ES (M = 261.06) each had a mean reading scale score that was significantly greater than Jeffers ES, Manch ES, and Petersen ES with Cambeiro ES (M = 264.28).
Cortez ES (M = 261.64), Moore ES (M = 260.79), Rundle ES (M = 262.57), Thomas ES (M = 262.11), Ullom ES (M = 263.35), Warren ES (M = 260.86), and Wengert ES (M = 263.16) all significantly greater than Jeffers ES and Petersen ES. And finally, McCall ES had a mean reading scale score (M = 262.25) that was significantly greater than Jeffers ES.

An examination of those schools with mean reading scale scores significantly lower than expected revealed Ronnow ES had a scale score significantly greater than Jeffers ES, Manch ES, and Peterson ES, while at the same time significantly less than Harmon ES. This pattern was repeated at Lake ES and Hewetson ES, with both significantly greater than Jeffers ES and Peterson ES, while at the same time significantly less than Harmon ES, with Tom Williams ES having a mean reading scale score that was significantly greater than Jeffers ES and at the same time significantly less than Bracken ES, Crestwood ES, Harmon ES, and Twin Lakes ES.

Dearing ES had a mean reading scale score that was significantly less than Harmon ES only with Craig ES, Detwiler ES, and Sunrise Acres ES all significantly less than Bracken ES, Crestwood ES, Harmon ES and Twin Lakes ES. Wynn ES demonstrated measurements similar to Craig ES, Detwiler ES, and Sunrise Acres ES with the addition of having a mean reading scale score that was significantly lower than Lincoln ES. Ira Earl ES, Hollingsworth ES, and Long ES were each similar to Wynn ES with an addition significantly lower relationship to Red Rock ES while Tate ES and Lunt ES each had reading scale scores significantly less than Bracken ES, Crestwood ES, Edwards ES, Harmon ES, Lincoln ES, Red Rock ES and Twin Lakes ES. Culley ES was similar to Tate ES and Lunt ES with an additional significantly lower mean reading scale
score relationship to Vegas Verdes ES. Bell ES was similar to Culley ES as well as significantly lower than Snyder ES and Squires ES while Herron ES continued this pattern with a profile similar to Bell ES as well as being significantly lower than Dailey ES. And finally, Manch ES was similar to Bell ES with the inclusion of being significantly lower than Gragson ES, Lynch ES, Park ES, Ronnow ES and Woolley ES with Jeffers ES demonstrating a mean reading scale score significantly less than all schools in the Hispanic subgroup with the exception of Bell ES, Craig ES, Culley ES, Dearing ES, Detwiler ES, Ira Earl ES, Herron ES, Hollingsworth ES, Jeffers ES, Long ES, Lunt ES, Manch ES, Petersen ES, Pittman ES, Ronzone ES, Rowe ES, Sunrise Acres ES, Tate ES, Mountain View ES, and Wynn ES, while Petersen ES was similar to Jeffers ES less a significantly lower relationship to McCall ES and Tom Williams ES. As a final note, Pittman ES, Rowe ES, Ronzone ES and Mountain View ES demonstrated mean reading scale scores with no significant relationships to any of the schools in the Hispanic cluster.

An examination of the math mean scale score ANOVA procedure for the Hispanic cluster, less outliers Cahlan ES and Petersen ES with the Tamhane’s T2 post-hoc procedure applied, resulted in Twin Lakes ES (M = 305.25) and Harmon ES (M = 303.00) each having mean math scale scores significantly greater than Bell ES (M = 254.74), Cortez ES (M = 273.92), Culley ES (M = 256.96), Dearing ES (M = 264.31), Detwiler ES (M = 270.54), Ira Earl ES (M = 274.72), Edwards ES (M = 269.94), Herron ES (M = 266.70), Hollingsworth ES (M = 252.35), Jeffers ES (M = 249.37), Lake ES (M = 271.99), Long ES (M = 272.13), Lunt ES (M = 260.70), Manch ES (M = 257.35), Moore ES (M = 270.51), Ronzone ES (M = 271.11), Rowe ES (M = 273.92), Tate ES (M
Mountain View ES (M = 268.18), Wengert ES (M = 271.85), Tom Williams ES (M = 277.87), Woolley ES (M = 271.80), and Wynn ES (M = 253.36). A similar relationship was found with Gragson ES (M = 296.93), which had a mean math scale score that was significantly greater than Bell ES, Culley ES, Dearing ES, Detwiler ES, Ira Earl ES, Edwards ES, Herron ES, Hollingsworth ES, Jeffers ES, Lake ES, Long ES, Lunt ES, Manch ES, Moore ES, Ronzone ES, Woolley ES, and Wynn ES, while the mean math scale score at Vegas Verdes ES (M = 298.32) was significantly greater than Bell ES, Culley ES, Dearing ES, Edwards ES, Herron ES, Hollingsworth ES, Jeffers ES, Lunt ES, Manch ES, Moore ES, Woolley ES, and Wynn ES. Crestwood ES (M = 296.79) had a mean math scale score that was similar to Vegas Verdes ES less the significantly greater relationship with Moore ES followed by Red Rock ES (M = 297.19), also similar to Vegas Verdes ES less the significantly greater relationship with Moore ES and Woolley ES. Lincoln ES (M = 292.12) had a mean math scale score that was significantly greater than Bell ES, Culley ES, Herron ES, Hollingsworth ES, Jeffers ES, Lunt ES, Manch ES, and Wynn ES, with Craig ES (M = 289.65), Ronnow ES (M = 286.91) and Thomas ES (M = 290.21) all similar to Lincoln ES with the exception of no significant relationship to Herron ES.

Additional Hispanic cluster schools having mean reading scale scores that were significantly greater than expected were Dailey ES (M = 285.13), Hewetson ES (M = 285.68) and Rundle ES (M = 286.898), each significantly greater than Bell ES, Culley ES, Hollingsworth ES, Jeffers ES, Manch ES, and Wynn ES, with Cambeiro ES (M = 285.67), Lynch ES (M = 283.54), and Snyder ES (M = 285.12) similar to Dailey ES, Hewetson ES, and Rundle ES less the significantly greater relationship to Culley ES.
Squires ES (M = 279.98) had a mean reading scale score that was significantly greater than Bell ES, Hollingsworth ES, Jeffers ES, and Wynn ES, with Park ES (M = 282.26) significantly greater than Bell ES, Hollingsworth ES, and Jeffers ES, while Warren ES (M = 282.03) was significantly greater than Hollingsworth ES and Jeffers ES. And finally, Bracken ES (M = 280.71), McCall ES (M = 279.67), Sunrise Acres ES (M = 276.62), and Ullom ES (277.34) each had mean reading scale scores that were significantly greater than Jeffers ES.

Examining those schools in the Hispanic cluster with lower than expected mean scale scores revealed Tom Williams ES had a mean math scale score that was significantly greater than Bell ES, Hollingsworth ES, and Jeffers ES, while at the same time significantly lower than Harmon ES and Twin Lakes ES. Ira Earl ES demonstrated a similar characteristic by having a mean math scale score that was significantly greater than Jeffers ES while at the same time significantly lower than Gragson ES, Harmon ES, and Twin Lakes ES, while Woolley ES had a mean math scale score that was also significantly greater than Jeffers ES while at the same time significantly less than Crestwood ES, Gragson ES, Harmon ES, Twin Lakes ES and Vegas Verdes ES. Cortez ES, Mountain View ES, Rowe ES, Tate ES, and Wengert ES were all significantly less than Harmon ES and Twin Lakes ES, with Detwiler ES, Lake ES, Long ES, and Ronzone ES significantly less than Gragson ES, Harmon ES and Twin Lakes ES, while Moore ES had a mean math scale score that was significantly less than Gragson ES, Harmon ES, Twin Lakes ES, and Vegas Verdes ES. Dearing ES and Edwards ES were each significantly less than Crestwood ES, Gragson ES, Harmon ES, Red Rock ES, Twin Lakes ES, and Vegas Verdes ES, with Herron ES significantly lower than Crestwood ES.
Gragson ES, Harmon ES, Lincoln ES, Red Rock ES, Twin Lakes ES, and Vegas Verdes ES. Lunt ES had a mean math scale score that was similar in significance to Herron ES as well as significantly lower than Craig ES, Ronnow ES, and Thomas ES, with Culley ES similar to Lunt ES as well as being significantly lower than Dailey ES, Hewetson ES, and Rundle ES, while Manch ES was similar to Culley ES with the addition of being significantly lower than Cambeiro ES, Lynch ES, and Snyder ES. Wynn ES had a mean math scale score that was significantly lower than Cambeiro ES, Craig ES, Crestwood ES, Dailey ES, Gragson ES, Harmon ES, Hewetson ES, Lincoln ES, Lynch ES, Red Rock ES, Ronnow ES, Rundle ES, Snyder ES, Squires ES, Thomas ES, Twin Lakes ES, and Vegas Verdes ES with Bell ES similar to Wynn ES, as well as being significantly lower than Park ES and Tom Williams, while Hollingsworth ES was similar to Bell ES, with the inclusion of a significantly lower relationship to Warren ES. And finally, Jeffers ES had a mean math scale score that was significantly lower than Bracken ES, Cambeiro ES, Craig ES, Crestwood ES, Dailey ES, Ira Earl ES, Gragson ES, Harmon ES, Hewetson ES, Lincoln ES, Lynch ES, McCall ES, Park ES, Red Rock ES, Ronnow ES, Rundle ES, Snyder ES, Squires ES, Sunrise Acres ES, Thomas ES, Twin Lakes ES, Ullom ES, Vegas Verdes ES, Warren ES, T. Williams ES, and Woolley ES. Pittman ES (M = 273.17) was the only school in the Hispanic cluster that did not demonstrate significant mean math scale score relationships.

In summary, Calhan ES was removed from the Hispanic cluster reading analysis as an outlier having above expected values with Bracken ES, Crestwood ES, Harmon ES, and Twin Lakes ES each demonstrating mean reading scale scores that were significantly greater than the mean reading performance of Jeffers ES, Manch ES, and Peterson ES.
The remaining schools demonstrated significance in mean reading scale scores to a lesser degree with Pittman ES, Ronzone ES, Rowe ES, and Mountain View ES providing no significant relationships. The analysis of Hispanic math mean scale scores began with Calhan ES removed as an outlier above expected distributions with Petersen ES removed as an outlier below expected distributions, leaving Gragson ES, Harmon ES and Twin Lakes ES reporting mean math scale scores that were significantly greater when compared to Bell ES, Jeffers ES, Hollingsworth ES, Manch ES and Wynn ES. Similar to the reading mean scale score analysis, many Hispanic schools demonstrated significant relationships in mean math scale score to a lesser degree with Pittman ES again providing no significant relationships.

**White Subgroup Analysis**

The White cluster reading median scale score of 324.06 with Q1 of 312.21 and Q3 of 335.44 resulted in an IQR of 23.23. An evaluation of outliers using $1.5 \times IQR$ (34.85) resulted in extreme values of 277.36 and 370.89, eliminating all White cluster schools as potential reading outliers. Repeating this procedure, using the white cluster mean math scale scores, resulted in a median scale score of 339.62, Q1 of 328.55, Q3 of 351.07, IQR of 22.52, and $1.5 \times IQR$ of 33.78. Applying the assessment to Q1 and Q3 resulted in extreme values of 294.77 and 384.85, again outside the range of white mean math scale scores. Based upon these results, all schools in the White cluster were found within the outlier extreme values for both reading and math, allowing the entire cluster of schools to be included in the ANOVA procedure.
The main ANOVA summary table for the White cluster, found in table 23, reported a reading $F(28,7213) = 11.18, p < .05,$ and a math $F(29,7213) = 13.91, p < .05,$ indicating a significant difference in mean reading and mean math scale scores did exist in at least one pair of schools for each respective subject area. As Levene's test of homogeneity of variances illustrated in table 24, the test of equality of variances among schools in the White cluster revealed significant differences in reading scale score variances, $F(29,7213) = 2.02, p < .05,$ with math non-significant, $F(29,7213)=1.21, p > .05$. With the assumption of equal variances in math, standard ANOVA procedures could be carried out using Tukey's HSD post-hoc analysis. With the assumption of homogeneity of variances in reading violated, Welch's robust test of equality of means was reported with post-hoc analysis carried out using Tamhane's T2 procedure. Welch's robust test of equality, found in table 25, reported a significant reading relationship with $F(29,2388.42)=11.02, p < .05,$ meaning the ANOVA procedure could be carried out for reading using Tamhane's T2 post-hoc procedure. The White cluster reading and math mean scale scores, standard deviations, standard errors, 95% confidence intervals, and minimum/maximum values can be found in Appendix F with post-hoc table of differences and significance levels in Appendix G.
Table 23. **White ANOVA Summary Table**

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS Between Groups</td>
<td>1491698.52</td>
<td>29</td>
<td>51437.88</td>
<td>11.18</td>
<td>0.00</td>
</tr>
<tr>
<td>Within Groups</td>
<td>33193165.79</td>
<td>7213</td>
<td>4601.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>34684864.31</td>
<td>7242</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSS Between Groups</td>
<td>2021323.03</td>
<td>29</td>
<td>69700.79</td>
<td>13.91</td>
<td>0.00</td>
</tr>
<tr>
<td>Within Groups</td>
<td>36153118.60</td>
<td>7213</td>
<td>5012.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>38174441.63</td>
<td>7242</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 24. **Levene's Test of Homogeneity of Variance for White**

<table>
<thead>
<tr>
<th></th>
<th>Levene Statistic</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS</td>
<td>2.02</td>
<td>29</td>
<td>7213</td>
<td>0.00</td>
</tr>
<tr>
<td>MSS</td>
<td>1.21</td>
<td>29</td>
<td>7213</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Table 25. **Welch Robust Tests of Equality of Means for White**

<table>
<thead>
<tr>
<th></th>
<th>Statistic</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS</td>
<td>Welch</td>
<td>11.02</td>
<td>29</td>
<td>2388.92</td>
</tr>
<tr>
<td>MSS</td>
<td>Welch</td>
<td>13.82</td>
<td>29</td>
<td>2387.90</td>
</tr>
</tbody>
</table>

Twitchell ES was found to have the highest mean reading scale score in the White cluster, M = 356.43, which was also significantly greater than Bonner ES (M = 323.65), Richard Bryan ES (M = 330.91), Conners ES (M = 324.47), David Cox ES (M = 328.89), Darnell ES (M = 308.91), Derfelt ES (M = 320.07), Eisenberg ES (M = 312.26), Galloway ES (M = 311.97), Gibson ES (M = 327.66), Hill ES (M = 312.95), Jacobson...
ES (M = 314.99), May ES (M = 320.47), Morrow ES (M = 333.55), Newton ES (M = 317.24), Parson ES (M = 311.50), Piggott ES (M = 326.53), Rhodes ES (M = 312.21), Glen Taylor ES (M = 310.80), Tobler ES (M = 305.41), Tomiyasu ES (M = 305.08), and Walker ES (M = 312.06). Lamping ES (M = 345.82), Ober ES (M = 347.53), and Staton ES (M = 348.81) each had mean reading scale scores that were significantly greater than Bonner ES, Darnell ES, Derfelt ES, Eisenberg ES, Galloway ES, Hill ES, Jacobson ES, May ES, Newton ES, Parson ES, Rhodes ES, Glen Taylor ES, Tobler ES, Tomiyasu ES, and Walker ES, with Lummis ES (M = 341.05) and Vanderburg ES (M = 340.22) significantly greater than Darnell ES, Eisenberg ES, Galloway ES, Hill ES, Jacobson ES, Newton ES, Parson ES, Rhodes ES, Glen Taylor ES, Tobler ES, Tomiyasu ES, and Walker ES. Scherkenbach ES had a mean reading scale score (M = 335.44) significantly greater than Darnell ES, Eisenberg ES, Galloway ES, Hill ES, Rhodes ES, Glen Taylor ES, Tobler ES, Tomiyasu ES, and Walker ES, with Bilbray ES (M = 335.29) significantly greater than Darnell ES, Galloway ES, Tobler ES, Tomiyasu ES, and Walker ES. The final schools to demonstrate significantly greater mean reading scale scores were McDoniel ES (M = 336.40), which was significantly greater than Darnell ES and Tobler ES with Morrow ES significantly greater than Darnell ES, Galloway ES, Tobler ES, Tomiyasu ES, and Walker ES, while at the same time significantly less than Twitchell ES. Richard Bryan ES, Conners ES, David Cox ES, Gibson ES, and Piggott ES each had mean reading scale scores that were significantly lower than Twitchell ES with Bonner ES, Derfelt ES, and May ES each significantly lower than Lamping ES, Ober ES, Staton ES, and Twitchell ES. Jacobson ES, Newton ES, and Parson ES each had reading scale scores significantly lower than Lamping ES, Lummis ES, Ober ES,

The school with the greatest mean math scale score was again Twitchell ES, having a mean scale score of 374.27 and significantly greater than Bilbray ES (M = 341.70), Bonner ES (M = 351.07), Richard Bryan ES (M = 346.96), Conners ES (M = 328.55), Darnell ES (M = 311.51), Derfelt ES (M = 330.31), Eisenberg ES (M = 334.75), Galloway ES (M = 322.03), Gibson ES (M = 347.72), Hill ES (M = 326.49), Jacobson ES (M = 333.06), Lummis ES (M = 344.01), May ES (M = 338.05), Morrow ES (M = 336.70), Newton ES (M = 339.18), Parson ES (M = 316.15), Piggott ES (M = 340.06), Rhodes ES (M = 322.70), Scherkenbach ES (M = 342.73), Glen Taylor ES (M = 346.45), Tobler ES (M = 307.58), Tomiyasu ES (M = 321.51), and Walker ES (M = 335.82). This was followed by Lamping ES (M = 363.82) and Staton ES (M = 364.77), each having mean math scale scores significantly greater than Bilbray ES, Conners ES, Darnell ES, Derfelt ES, Eisenberg ES, Galloway ES, Hill ES, Jacobson ES, May ES, Morrow ES, Newton ES, Parson ES, Rhodes ES, Tobler ES, Tomiyasu ES, and Walker ES, with Ober ES (M = 362.72) and Vanderburg ES (M = 361.70) also significantly

In an interesting shift in the White cluster measures of significance, Bonner ES had a mean math scale score that was significantly less than Twitchell ES and significantly greater than Darnell ES, Galloway ES, Hill ES, Parson ES, Rhodes ES, Tobler ES, and Tomiyasu ES. This pattern continued with Gibson ES and Glen Taylor ES, each significantly less than Twitchell ES while at the same time significantly greater than Darnell ES, Galloway ES, Parson ES, Rhodes ES, and Tobler ES, followed by Richard Bryan ES, which was significantly less than Twitchell ES and significantly greater than Darnell ES, Galloway ES, Parson ES, and Tobler ES with Lummis ES, Piggott ES, and Scherkenbach all significantly less than Twitchell ES and significantly greater than Darnell ES and Tobler ES.

Bilbray ES and Newton ES each had mean math scale scores that were significantly less than Lamping ES, Staton ES, and Twitchell ES while at the same time significantly greater than Darnell ES and Tobler ES, with May ES significantly less than Lamping ES, Ober ES, Staton ES, Twitchell ES, and Vanderburg ES while significantly greater than Darnell ES and Tobler ES, followed by Morrow ES and Walker ES, each significantly lower than Lamping ES, McDoniel ES, Ober ES, Staton ES, Twitchell ES,
and Vanderburg ES while at the same time significantly greater than Darnell ES and Tobler ES.

The first school to provide a mean math scale score with no positive levels of significance was Eisenberg ES, which was significantly less than Lamping ES, Ober ES, Staton ES, Twitchell ES, and Vanderburg ES. This continued with Conners ES, Derfelt ES, and Jacobson ES, each having mean math scale scores that were significantly less than Lamping ES, McDoniel ES, Ober ES, Staton ES, Twitchell ES, and Vanderburg ES, with Tomiyasu ES and Hill ES significantly lower than Bonner ES, David Cox ES, Lamping ES, McDoniel ES, Ober ES, Staton ES, Twitchell ES, and Vanderburg ES. Rhodes ES had a mean math scale score that was significantly less than Bonner ES, David Cox ES, Gibson ES, Lamping ES, McDoniel ES, Ober ES, Staton ES, Glen Taylor ES, Twitchell ES, and Vanderburg ES, with Galloway ES and Parson ES significantly less than Bonner ES, Richard Bryan ES, David Cox ES, Gibson ES, Lamping ES, McDoniel ES, Ober ES, Staton ES, Glen Taylor ES, Twitchell ES, and Vanderburg ES.

And finally, Darnell ES and Tobler ES had mean math scale scores that were significantly less than Bilbray ES, Bonner ES, Richard Bryan ES, David Cox ES, Gibson ES, Lamping ES, Lummis ES, May ES, McDoniel ES, Morrow ES, Newton ES, Ober ES, Piggott ES, Scherkenbach ES, Staton ES, Glen Taylor ES, Twitchell ES, Walker ES, and Vanderburg ES.

To summarize, Lamping ES, Lummis ES, Ober ES, Staton ES, Twitchell ES, and Vanderburg ES all had mean reading scale scores that were significantly greater than expected with Darnell ES, Galloway ES, Tobler ES, Tomiyasu ES, and Walker ES significantly lower than expected. A similar pattern was found during the mean math
scale score ANOVA procedure, with Lamping ES, McDaniels ES, Ober ES, Staton ES, Twitchell ES, and Vanderburg ES all providing mean math scale scores significantly greater than expected while Darnell ES, Parson ES, and Tobler ES all providing mean math scale scores significantly lower than expected. The remaining schools varied above or below average mean reading or mean math scale scores to a somewhat lesser degree of significance with no schools identified as outliers.

**IEP Subgroup Analysis**

The final cluster group identified for analysis was the 35 schools identified as having larger than expected populations of students with individualized education plans (IEP). A median reading scale score of 229.77, Q1 of 212.50, Q3 of 257.41, IQR of 44.91, and 1.5 x IQR of 67.37 resulted in extreme values of 145.03 and 324.78. Considering the minimum mean reading scale score for the IEP cluster was 168.36 (Griffith ES) and maximum mean reading scale score for the IEP cluster was 288.49 (Heckethorn ES), no schools were identified as having outlier mean reading scale scores. Repeating the IEP outlier procedure for mean math scale scores resulted in a median value of 251.74, Q1 of 229.36, Q3 of 274.32, IQR of 44.96, 1.5 x IQR of 67.44, and extreme values of 161.62 and 341.76. With a minimum mean math scale score of 187.04 (Griffith ES) and maximum mean math scale score of 307.91 (Dooley ES) no schools were identified as having outlier mean math scale scores. Therefore, all schools identified as having larger than expected populations of IEP students were well within the extreme values necessary for identification as outliers in reading or math, requiring all IEP cluster schools to be included in the ANOVA procedure.
The main IEP ANOVA summary table, found in table 26, identified a significant difference in mean reading scale scores between at least one pair of schools, $F(34,1551) = 6.96$, $p < .05$. This significant relationship was also found when comparing across school mean math scale scores, with math $F(34,1551) = 6.77$, $p < .05$. The IEP ANOVA post-hoc analysis for both reading and math was carried out using Tamhane’s T2 procedure, with Levene’s test for homogeneity of variances significant in both reading, $F(34.1551) = 2.10$, $p < .05$, and math, $F(34.1551) = 1.78$, $p < .05$ (table 27), while Welch’s robust tests of equality of means significant in reading, $F(34,520.83) = 8.82$, $p < .05$, and math $F(34,520.65) = 7.37$, $p < .05$ (table 28). The IEP cluster reading and math mean scale scores, standard deviations, standard errors, 95% confidence intervals, and minimum/maximum values can be found in Appendix F with post-hoc table of differences and significance levels in Appendix G.

<table>
<thead>
<tr>
<th>Table 26. <strong>IEP ANOVA Summary Table</strong></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>1417525.36</td>
<td>34</td>
<td>41691.92</td>
<td>6.96</td>
<td>0.00</td>
</tr>
<tr>
<td>Within Groups</td>
<td>9293777.36</td>
<td>1551</td>
<td>5992.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>10711302.72</td>
<td>1585</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>1477879.99</td>
<td>34</td>
<td>43467.06</td>
<td>6.77</td>
<td>0.00</td>
</tr>
<tr>
<td>Within Groups</td>
<td>9960835.76</td>
<td>1551</td>
<td>6422.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11438715.75</td>
<td>1585</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 27. **Levene's Test of Homogeneity of Variance for IEP**

<table>
<thead>
<tr>
<th>Levene Statistic</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS</td>
<td>2.10</td>
<td>34</td>
<td>1551</td>
</tr>
<tr>
<td>MSS</td>
<td>1.78</td>
<td>34</td>
<td>1551</td>
</tr>
</tbody>
</table>

Table 28. **Welch Robust Tests of Equality of Means for IEP**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS</td>
<td>8.82</td>
<td>34</td>
<td>520.83</td>
</tr>
<tr>
<td>MSS</td>
<td>7.37</td>
<td>34</td>
<td>520.65</td>
</tr>
</tbody>
</table>

Within the cluster of schools identified as having larger than expected populations of IEP students, Heckethorn ES (M = 288.49) had the largest mean reading scale score, which was significantly greater than Antonello ES (M = 216.28), Beckley ES (M = 217.27), Cozine ES (M = 229.77), Fyfe ES (M = 197.14), Griffith ES (M = 168.36), Heard ES (M = 212.50), Hickey ES (M = 198.82), Katz ES (M = 215.28), Martin L. King ES (M = 213.46), Martinez ES (M = 176.40), McCaw ES (M = 215.32), McWilliams ES (M = 175.62), Hal Smith ES (M = 206.41), Sewell ES (M = 210.41), Stanford ES (M = 205.15), and Robert Taylor ES (M = 212.91). This was followed by Roberts ES (M = 272.48), which was significantly greater than Fyfe ES, Griffith ES, Hickey ES, Martin L. King ES, Martinez ES, McCaw ES, McWilliams ES, Hal Smith ES, and Stanford ES, with Kahre ES (M = 272.30) and Helen Smith ES (M = 271.36) both significantly greater than Fyfe ES, Griffith ES, Hickey ES, Martinez ES, McWilliams ES, Hal Smith ES, and Stanford ES. Bartlett ES (M = 263.10) had a mean reading scale score that was significantly greater than Fyfe ES, Griffith ES, Martinez ES, and McWilliams, with
Adcock ES (M = 248.36), Allen ES (M = 264.09), Carl ES (M = 238.04) Christensen ES (M = 251.00), Dondero ES (M = 245.00), Dooley ES (M = 262.47), Garehime ES (M = 257.41), Kesterson ES (M = 259.36), Neal ES (M = 252.21), and Tarr ES (M = 244.42) all significantly greater than Griffith ES, Martinez ES, and McWilliams ES. Ferron ES (M = 234.68) had a mean reading scale score that was significantly greater than Griffith ES and Martinez ES, with Cozine ES significantly greater than Griffith ES, Martinez ES, and McWilliams ES while significantly lower than Heckethorn ES.

A summary of schools at the lower end of the reading scale score distribution revealed Antonello ES, Beckley ES, Heard ES, Katz ES, Sewell ES, and Robert Taylor ES all had reading scale scores that were significantly lower than Heckethorn ES, with Martin L King ES and McCaw ES both significantly lower than Heckethorn ES and Roberts ES. Hickey ES, Hal Smith ES, and Stanford ES all had reading scale scores that were significantly lower than Heckethorn ES, Kahre ES, Roberts ES, and Helen Smith ES, with Fyfe ES significantly lower than Bartlett ES, Heckethorn ES, Kahre ES, Roberts ES, and Helen Smith ES. And finally, McWilliams ES had a mean reading scale score that was significantly lower than Adcock ES, Allen ES, Bartlett ES, Carl ES, Christensen ES, Cozine ES, Dondero ES, Garehime ES, Heckethorn ES, Kahre ES, Kesterson ES, Neal ES, Roberts ES, Helen Smith ES, and Tarr ES, with Griffith ES and Martinez ES both significantly lower than Adcock ES, Allen ES, Bartlett ES, Carl ES, Christensen ES, Cozine ES, Dondero ES, Ferron ES, Garehime ES, Heckethorn ES, Kahre ES, Kesterson ES, Neal ES, Roberts ES, Helen Smith ES, and Tarr ES. Deskin ES (M = 229.25), Hinman ES (M = 222.96), and Mack ES (M = 229.82) all demonstrated no statistically
significant relationships in mean reading scale scores when compared to schools identified in the IEP cluster.

The IEP cluster school demonstrating the highest mean math scale score was Dooley ES (M = 307.91), which was significantly greater than Beckley ES (M = 233.56), Carl ES (M = 237.87), Cozine ES (M = 251.74), Fyfe ES (M = 233.66), Griffith ES (M = 187.04), Heard ES (M = 219.92), Hickey ES (M = 218.76), Hinman ES (M = 233.62), Martinez ES (M = 201.40), McCaw ES (M = 207.60), McWilliams ES (M = 188.15), Sewell ES (M = 229.36), Hal Smith ES (M = 221.39), and Stanford ES (M = 213.40). This was followed by Kahre ES (M = 291.85), which had a mean math scale score that was significantly greater than Beckley ES, Griffith ES, Hickey ES, Martinez ES, McCaw ES, McWilliams ES, Hal Smith ES, and Stanford. Garehime ES (M = 288.17), Heckethorn ES (M = 291.75), and Roberts ES (M = 283.79) were all significantly greater than Griffith ES, Hickey ES, Martinez ES, McCaw ES, McWilliams ES, Hal Smith ES, and Stanford ES, with Bartlett ES (M = 285.41) significantly greater than Griffith ES, Martinez ES, McCaw ES, McWilliams ES, Hal Smith ES, and Stanford. Helen Smith ES (M = 291.36) and Allen ES (M = 280.66) had mean math scale scores that were significantly greater than Griffith ES, Martinez ES, McCaw ES, and McWilliams ES, with Adcock ES (M = 271.45), Christensen ES (M = 265.48), Dondono ES (M = 271.58), and Kesterson ES (M = 265.17) all significantly greater Griffith ES, Martinez ES, and McWilliams ES. And finally, Antonello ES (M = 260.30) and Deskin ES (M = 274.32) had mean math scale scores that were significantly greater than Griffith ES and McWilliams ES with Ferron ES (M = 251.68), Mack ES (M = 260.66), and Tarr ES (M = 260.24) significantly greater than McWilliams ES.
Cozine ES had a mean math scale score that was significantly greater than McWilliams ES and significantly less than Dooley ES, with Carl ES, Fyfe ES, Heard ES, Hinman ES and Sewell ES all having mean math scale scores that were significantly less than Dooley ES. Beckley ES had a mean math scale score that was significantly lower than Kahre ES and Dooley ES with Hickey ES significantly lower than Dooley ES, Garehime ES, Heckethorn ES, Kahre ES, and Roberts ES. Hal Smith ES and Stanford ES were significantly less than Bartlett ES, Dooley ES, Garehime ES, Heckethorn ES, Kahre ES, and Roberts ES, with McCaw ES significantly less than Allen ES, Bartlett ES, Dooley ES, Garehime ES, Heckethorn ES, Kahre ES, Roberts ES, and Helen Smith ES, while Martinez ES was significantly less than Adcock ES, Allen ES, Bartlett ES, Christensen ES, Dondero ES, Dooley ES, Garehime ES, Heckethorn ES, Kahre ES, Kesterson ES, Roberts ES, and Helen Smith ES. And finally, Griffith ES had a mean math scale score that was significantly less than Adcock ES, Allen ES, Antonello ES, Bartlett ES, Christensen ES, Deskin ES, Dondero ES, Dooley ES, Garehime ES, Heckethorn ES, Kahre ES, Kesterson ES, Roberts ES, and Helen Smith ES, with McWilliams ES significantly less than Adcock ES, Allen ES, Antonello ES, Bartlett ES, Christensen ES, Cozine ES, Deskin ES, Dondero ES, Dooley ES, Ferron ES, Garehime ES, Heckethorn ES, Kahre ES, Kesterson ES, Mack ES, Roberts ES, Helen Smith ES and Tarr ES. Katz ES (M = 245.31), Martin L. King Jr. ES (M = 244.63), Neal ES (M = 253.87), and Robert Taylor ES (M = 233.60) demonstrated no statistically significant relationships with respect to mean math scale score relative to schools identified in the IEP cluster.
In summary, no schools identified as outliers in the IEP cluster. Heckethorn ES, Roberts ES, Kahre ES, and Helen Smith ES were identified as having mean reading scale scores significantly greater than expected with Griffith ES, Martinez ES, and McWilliams ES identified as having mean reading scale scores significantly lower than expected. Dooley ES, Kahre ES, Heckethorn ES, Roberts ES, Garehime ES, and Bartlett ES demonstrated IEP mean math scale scores significantly greater than anticipated with McWilliams ES, Griffith ES, and Martinez ES significantly lower than expected. And finally, the mean reading scale scores for Deskin ES, Hinman ES, and Mack ES were statistically insignificant with the mean math scale scores for Katz ES, Martin L. King Jr. ES, Neal ES, and Robert Taylor ES statistically insignificant.

Summary

Using student demographic data from the 2005-2006 Nevada standards based criterion referenced test, five unique, homogeneous clusters of schools were formed representing schools with larger than average populations from the subgroups African American, Asian, Hispanic, White, and IEP. Using standard ANOVA procedures, significance testing identified schools in each respective cluster having significantly higher or significantly lower mean reading or mean math scale scores. Standard ANOVA procedures determined if a significant difference in mean reading or mean math scale scores existed, with Levene’s test of homogeneity of variances or Welch’s robust tests of equality of means included. If the standard ANOVA procedure found a significant difference in mean reading or mean math scale score (P < .05), and Levene’s test concluded variances were equal (P > .05), post-hoc analysis was performed using
Tukey’s HSD. If Levene’s test identified unequal variances (P < .05), the Welch robust tests of equality of means was used as an alternative. In such cases, if the standard ANOVA procedure was significant (P < .05), and Welch’s test identified unequal means (P < .05), post-hoc analysis was carried out using Tamhane’s T2 procedure.

All final clusters (African American, Asian, Hispanic, White, and IEP) were found to have significant within group differences in both mean reading and mean math scale scores using the standard ANOVA procedures (P < .05). Equal variances, requiring Tukey’s HSD post-hoc procedure, were found in the Asian mean reading scale scores as well as the Asian and White mean math scale scores. Unequal variances, requiring Tamhane’s T2 post-hoc procedure, were found with the African American, Hispanic, White, and IEP mean reading scale scores as well as the African American, Hispanic, and IEP mean math scale scores.

Hoggard ES, Booker ES, and Mackey ES were identified as having mean reading scale scores that were significantly greater than most schools in the African American cluster with Tartan ES, Elizondo ES, Fitzgerald ES, and Kelly ES significantly lower. The remaining African American schools, which included Carson ES, Gilbert ES, Priest ES, Reed ES, Watson ES, Wilhelm ES, and W. Williams ES, demonstrated significance in mean reading scale score to a lesser degree. Kelly ES and Fitzgerald ES were identified as math outlier schools, having mean math scale scores below expected distributions, and removed from the African American math analysis. Hoggard ES, with a mean math scale score significantly above expected distributions, was also identified as an outlier school and removed from the African American math analysis. After removal of outlier schools Kelly ES, Fitzgerald ES, Hoggard ES, Gilbert ES and Mackey ES were
identified as having mean math scale scores significantly greater than Tartan ES. All remaining African American schools demonstrated no significant relationships in mean math scale scores.

Thiroit ES, having a mean reading and mean math scale score below expected distributions, was identified as an outlier school in the Asian cluster and removed from both reading and math analysis. Bendorf ES, with a mean math scale score above expected distributions, was identified as an outlier and removed from the Asian math analysis only. Bendorf ES, Goolsby ES, and Givens ES all demonstrated Asian cluster reading scale scores that were significantly greater with Givens ES, Rogers ES, Goolsby ES, Kim ES, and Frias ES demonstrating significantly greater mean math scale scores. Gehring ES and Diskin ES each had Asian cluster mean reading scale scores that were significantly lower than expected with Diskin ES, Hummel ES, and Jydstrup having significantly lower mean math scales scores. The mean reading and mean math scale scores for all remaining Asian cluster schools were, statistically speaking, equivalent.

In the Hispanic cluster, Calhan ES was identified as an outlier and removed from the reading and math analysis. Petersen ES, with a mean math scale score below expected distributions, was also identified as an outlier and removed from the Hispanic math analysis. As a result, Hispanic schools Bracken ES, Crestwood ES, Harmon ES, and Twin Lakes ES all demonstrated mean reading scale scores that were significantly greater with Jeffers ES, Manch ES, and Peterson ES significantly lower. Gragson ES, Harmon ES and Twin Lakes ES reporting math scale scores that were significantly greater while Bell ES, Jeffers ES, Hollingsworth ES, Manch ES and Wynn ES all reported mean math scale scores that were significantly low. The remaining Hispanic
schools demonstrated levels of significance in mean reading and mean math scale score
to a somewhat lesser degree with Ronzone ES, Rowe ES, and Mountain View ES
demonstrating no significant relationships in mean reading scale score while Pittman ES
demonstrating no significance in reading or math.

An analysis of those schools located in the White cluster found Lamping ES, Lummis ES, Ober ES, Staton ES, Twitchell ES, and Vanderburg ES having mean reading scale scores that were significantly greater than Darnell ES, Galloway ES, Tobler ES, Tomiyasu ES, and Walker ES. Lamping ES, McDoniel ES, Ober ES, Staton ES, Twitchell ES, and Vanderburg ES demonstrated mean math scale scores that were significantly above expectations with Darnell ES, Galloway ES, Parson ES, Rhodes ES, and Tobler ES demonstrating math scale scores below expectations. The remaining White cluster schools varied in significance above and below expectations to a lesser degree with no schools identified as outliers.

And finally, no schools from the IEP cluster were identified as outliers with Heckethorn ES, Roberts ES, Kahre ES, and Helen Smith ES all reporting reading scale scores significantly above expected distributions with Griffith ES, Martinez ES, and McWilliams ES reporting reading scale scores significantly below expectations. Dooley ES, Kahre ES, Heckethorn ES, Roberts ES, Garehime ES, and Bartlett ES demonstrated mean math scale scores significantly above expectations with McWilliams ES, Griffith ES, and Martinez ES reporting mean math scale scores below expectations. And finally, there were no significant relationships in the mean reading scale scores for IEP cluster schools Deskin ES, Hinman ES, and Mack ES with significant relationships in the mean
math scale scores for IEP cluster schools Katz ES, Martin L. King ES, Neal ES, and Robert Taylor ES.
CHAPTER 5

SUMMARY OF FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS FOR FURTHER STUDY

Introduction

It was the purpose of this study to examine the efficacy of NCLB when evaluating the overall academic performance and annual progress of Clark County School District elementary schools sharing unique demographic characteristics. It was hypothesized that the dominant student population would have a significant influence on academic performance, which if not considered may result in unnecessary sanctions with respect to NCLB legislation. Using an SPSS TwoStep cluster analysis procedure, schools were grouped based upon dominant student populations and determinations made concerning statistically significant differences in mean reading and mean math CRT scale scores for schools contained within homogeneous clusters.

Summary of Findings

The findings from this study will be organized and described around the three research questions.
Question 1: Which Demographic Variables, as Recognized Through NCLB, Generate Unique and Homogeneous Clusters of Five or More Schools

The initial TwoStep cluster analysis procedure generated eight unique clusters, representative of the eight subgroups as identified through NCLB, with a membership high of 30 schools for cluster 8 to a low of 12 schools for clusters 5 and 6. Considering the purpose of this study was to identify clusters of schools serving subgroup populations significantly greater than district averages, subgroup clusters must have exceeded the positive $t$ statistic at $\alpha = 0.05$, with Bonferroni adjustments applied, before statistical significance could be determined. Those subgroups identified as insignificant were eliminated from analysis.

The American Indian subgroup, with an average standardized value of 0.04, failed to demonstrate significant positive mean values and was therefore statistically insignificant and eliminated from analysis. Also, no elementary schools included in this study had an American Indian population large enough to satisfy the minimum requirements for inclusion in inference testing, further justifying eliminating from analysis the American Indian subgroup. The Asian subgroup generated an overall average value of 0.40 with cluster 6 positive and significant, indicating the 12 schools included in cluster 6 were representative of a group of schools having significantly larger populations of Asian students. An average value of 2.02 was recorded in the Hispanic subgroup with what appeared to be two unique and significant clusters, clusters 7 and 8, with both clusters positive and significant. The African American subgroup had an overall mean of 0.77, with cluster 4 representative of a single significant cluster.
Three significant cluster groups were identified with the white subgroup. With a mean value of 1.67, cluster 1, cluster 2, and cluster 3 were all unique and significant to the White subgroup while students with IEP’s demonstrated an overall mean value of 0.08 and were significant with respect to clusters 3 and cluster 5. The LEP subgroup had an overall mean value of 2.27 and significant to cluster 7 and cluster 8, which followed a pattern similar to the Hispanic subgroup. And finally, students identified as eligible for FRL had an overall mean value of 3.24 and were also significant with respect to cluster 7 and cluster 8.

The Pearson’s product-moment correlation coefficient verified a strong positive linear relationship existed between Hispanic, LEP, and FRL, with additional analysis confirming the Asian, African American, and White subgroups were unique and independent while Hispanic, LEP and FRL exhibited strong multicollinearity. The presence of multicollinearity between Hispanic, LEP, and FRL created redundant information, making it inappropriate to analyze the three subgroups separately. Therefore, further analysis eliminated the LEP and FRL subgroups from the data set with the Hispanic subgroup representative of the Hispanic, LEP, and FRL subgroups.

A second TwoStep cluster analysis procedure was performed with the American Indian subgroup removed from analysis due to insignificant populations and the Hispanic subgroup representative of Hispanic, LEP, and FRL. Phase II of the TwoStep cluster analysis procedure recognized the Asian subgroup, cluster 3, as a single, statistically significant cluster with a positive mean value of 3.19. The Hispanic subgroup, which was representative of Hispanic, LEP, and FRL, identified cluster 5 as a statistically significant cluster with a mean value of 10.02. Cluster 4 was statistically significant with
respect to the African American subgroup with a mean value of 10.70 while the White subgroup was identified by cluster 2 with a mean value of 8.71. And finally, the IEP subgroup was found to be statistically significant with respect to cluster 1 with a mean value of 0.88.

The initial TwoStep cluster analysis procedure, Phase I, served to identify insignificant variables and multicollinearity between subgroups while the second TwoStep cluster analysis procedure, Phase II, generated five unique and statistically significant clusters of schools, each containing five or more schools, using subgroup variables Asian, Hispanic, African American, White, and IEP. This would suggest that although NCLB requires Clark County School District to report the performance of the American Indian, Asian, Hispanic, African American, White, IEP, LEP, and FRL subgroups, statistically speaking the only subgroups capable of providing valid, non-redundant information would be the Asian, Hispanic, African American, White, and IEP subgroups. The American Indian, LEP, and FRL subgroups provided no additional information and simply increased the probability for not achieving adequate yearly progress.

Another interesting aspect was the comparison of performance using mean reading and mean math scale scores instead of targeted rates of proficiency. As discussed earlier, the baseline rate of proficiency was established by ranking schools in terms of subject area proficiency through cumulative enrollment up to and including the 20\textsuperscript{th} percentile. Baseline rates of proficiency, and future targeted rates of proficiency, were determined using the school-wide rate of proficiency for that school identified at the 20\textsuperscript{th} percentile. While this technique guaranteed 80\% of all school-wide measurements
achieved the targeted rate of proficiency that first measurement year, subgroup baseline performance was ignored, with a single subgroup not meeting the targeted rate of proficiency all that would be necessary for the entire school to not make AYP.

A tremendous amount of information is lost when considering rates of proficiency only. Using a scale score of 300, on a 100-500 range, as the single metric failed to recognize differences that were obvious when using mean reading and mean math scale scores. The comparison of mean reading and mean math scale scores within homogeneous clusters of schools, as demonstrated in this work, provided the accuracy necessary to determine levels of significance, which would be required to accurately determine which schools were progressing toward narrowing and possibly eliminating the achievement gap.

To provide valid results the number of samples is a major concern, especially in high stakes situations such as NCLB. Current Nevada policy requires that schools report the performance of subgroups with 25 or more students. The minimum n-count of 25 was included to insure previously ignored populations, such as minority and special needs populations, were not ignored, while at the same time providing valid results, as 25 is near the point where the $t$ or $z$ table stabilizes.

Evaluations using clusters of homogeneous schools, in most instances, eliminated the negative impact associated with small subgroup size. The African American cluster sampled a minimum of 66 students at Hoggard ES with a maximum of 160 students evaluated at Tartan ES. The Hispanic cluster sampled a minimum of 119 students at Mountain View ES with a maximum of 389 students sampled at Ronnow ES while the White cluster evaluated a minimum of 137 students at Parsons ES and a maximum of 381
students sampled at Walker ES. The Asian cluster achieved the minimum n-count for the state of Nevada, with a sample size of 25 students at Thiroit ES, while having a maximum Asian sample group of 199 at Alamo ES. The IEP subgroup was also near the minimum n-count for Nevada, with 26 students sampled at Hinman ES and a maximum IEP sample group of 72 students at Heckethorn ES. While the Asian and IEP subgroup minimum n-counts were near the minimum subgroup size for Nevada, they were nonetheless valid and therefore appropriate for reporting purposes.

It is important to note the cluster results found in this study were unique and significant to Clark County School District, with the potential for different results using a similar procedure with data from another school district. While small populations of American Indian students justified the exclusion of that subgroup from this study, another district may have a significant American Indian population and therefore require the subgroup be included. Also, due to large enrollments of poor, non-English speaking Hispanic populations, the multicollinearity found between Hispanic, LEP, and FRL might have been unique to Clark County School District elementary schools. While the focus of this study was to examine differences between urban elementary schools located in homogeneous clusters within the Clark County School District, it is important to note differences may also be found between districts, making each district and school unique dependent upon enrollment.
Question 2 and 3: Does a Statistically Significant Difference in Mean Reading or Mean Math CRT Scale Scores Exist Within Homogeneous Clusters of Five or More Schools?

The intention of this study was to cluster similar schools and, once clustered, determine if mean reading or mean math scale scores could be used to determine significant differences between schools located within homogeneous subsets. As has been demonstrated, schools within homogeneous clusters were identified as achieving above or below expectations in both reading and math, with levels of significance used to determine levels of achievement. Certain schools performed at extremes above or below expectations, requiring exclusion as outliers, while others demonstrated higher or lower levels of significance, with the majority achieving at or near average. It was found that clusters containing smaller numbers of schools provided results that were easily interpreted while clusters containing larger numbers of schools were quite complex, generating long lists of schools with statistically significant relationships, making it difficult to determine cutoff values with confidence.

To accomplish this task, schools at extreme levels of significance, both above and below, were identified as such and removed from the data set. Once removed, the assumption was made that any significance shared between eliminated school and remaining schools no longer existed. After all schools exhibiting extreme levels of significance were identified and removed, the remaining schools were assumed to be at or near cluster mean reading or mean math scale score expectations. Therefore, schools found to have mean reading or mean math scale scores significantly greater than subgroup expectations were identified as such and removed; those found to have mean
reading or mean math scale scores significantly less than subgroup expectations were identified as such and removed; with all remaining schools identified as demonstrating, statistically speaking, average mean reading or mean math scale scores.

African American Analysis

The African American subgroup formed a unique, homogeneous cluster of 14 schools contained in cluster 4, with a universal mean reading scale score of 256.62 and a universal mean math scale score of 259.51. Located in this cluster were Booker ES, Carson ES, Elizondo ES, Fitzgerald ES, Kelly ES, Gilbert ES, Hoggard ES, Mackey ES, Priest ES, Reed ES, Tartan ES, Watson ES, Wilhelm ES, and Wendell Williams ES.

While no schools were identified as having reading scale scores outliers, Kelly ES and Fitzgerald ES were identified as having math scale score outliers with mean math scale scores below expected distributions while Hoggard ES was identified as having a math scale score outlier above expected distributions. Therefore, for analysis purposes Kelly ES, Fitzgerald ES and Hoggard ES were removed from the math data set.

Had this type of analysis been used to evaluate the AYP academic performance for those schools identified in the African American cluster, Hoggard ES and Mackey ES would have been identified as achieving above expectations in reading and math with Booker ES identified as overachieving in reading only and Gilbert ES identified as overachieving in math only. Kelly ES and Fitzgerald ES would have been identified as underachieving in reading and math with Elizondo ES identified as underachieving in reading only and Tartan ES underachieving in Math only. All remaining schools in the African American cluster would have mean reading and mean math scale scores that
were statistically similar and therefore suitable for classification as making AYP, assuming all additional criteria, such as participation rate and average daily attendance, had been met.

**Asian Subgroup Analysis**

The Asian subgroup was identified as a unique, homogeneous cluster of 28 schools contained in cluster 3, with a universal mean reading scale score of 318.72 and a universal mean math scale score of 340.37. Located within this cluster of schools was Alamo ES, Bass ES, Batterman ES, Beatty ES, Bendorf ES, Roger Bryan ES, Cartwright ES, Decker ES, Diskin ES, Marion Earl ES, Frias ES, Gehring ES, Givens ES, Goolsby ES, Goynes ES, Hayes ES, Hummel ES, Iverson ES, Jydstrup ES, Kim ES, Mendoza ES, Ries ES, Rogers ES, Tanaka ES, Thiroit ES, Treem ES, Whitney ES, and Wolff ES. Thiroit ES was identified as an outlier in both reading and math with mean scale scores below expected distributions while Bendorf ES was identified as a math outlier with a math scale score above expected distributions. For analysis purposes Thiroit ES was removed from the reading and math data set with Bendorf ES removed from math data set only.

Again using levels of significance to evaluate adequate yearly progress, assuming all additional criteria had been satisfied, Bendorf ES, Givens ES, and Goolsby ES would have been identified as achieving above expectations in both reading and math with Rodgers achieving above expectations in math only. Thiroit ES and Diskin ES would have been identified as achieving below standards in both reading and math with Jydstrup
ES below standards in math only. All other schools in the Asian cluster would, statistically speaking, have similar reading and math scale scores.

**Hispanic Subgroup Analysis**


The large number of schools in the Hispanic cluster made it difficult to determine which schools were significantly above or below expected distributions. Cahan ES, identified as an outlier in both reading and math, was clearly performing above expectations. Also performing above expectations in both reading and math was Harmon.
ES, Crestwood ES, Twin Lakes ES, Lincoln ES, and Red Rock ES. Achieving above expectations in reading only was Bracken ES and Edwards ES with Gragson ES and Vegas Verdes ES achieving above expectations in math only.

Petersen ES was identified and removed as an outlier with mean reading and mean math scale scores below expected distributions. Also achieving below expectations in both reading and math was Jeffers ES, Manch ES, Bell ES, Culley ES, and Lunt ES. Achieving below expectations in reading only was Herron ES with Wynn ES and Hollingsworth ES achieving below expectations in math only. All remaining schools in the Hispanic cluster were, statistically speaking, similar.

As previous analysis has demonstrated, the Hispanic cluster is strongly correlated to, and therefore representative of, the LEP subgroup. The limitations introduced when testing non-English speaking students with an English-only assessment may call into question the reliability and validity of the mean reading and mean math scale scores for those schools identified in the Hispanic cluster. It is possible this construct irrelevant component may have distorted the results from a measure of reading or math abilities to a measure of English proficiency. Therefore, caution should be taken when interpreting results from the Hispanic cluster with further analysis into test reliability and validity suggested.

**White Subgroup Analysis**

The White subgroup was identified as a unique, homogeneous cluster of 30 schools contained in cluster 2, with a universal mean reading scale score of 325.91 and a universal mean math scale score of 341.04. Included in the White cluster was Bilbray

As with the Hispanic cluster, the large number of schools, along with the lack of outliers, made identification of significance difficult. What was apparent was Twitchell ES, Staton ES, Ober ES, Lamping ES and Vanderburg ES all demonstrated mean reading and mean math scale scores that were above expected values with Lummis ES and Scherkenbach ES achieving above expectations in reading only while McDoniel ES achieved above expectations in math only. Tobler ES, Darnell ES, Tomiyasu ES, and Galloway ES all performed below expectations in both reading and math with Walker ES performing below expectations in reading only and Parson ES, Rhodes ES, and Hill ES performing below expectations in math only. The remaining schools contained in the White cluster demonstrated mean reading and mean math scale scores that were statistically similar.

**IEP Subgroup Analysis**

The final cluster group was the 35 schools identified as enrolling larger than expected populations of students having individualized education plans (IEP). Included

From the 35 schools located in the IEP cluster, Heckethorn ES, Roberts ES, and Kahre ES demonstrated achievement above expectations in reading and math with Helen Smith ES achieving above expectations in reading only while Dooley ES and Garehime ES achieved above expectations in math only. The IEP cluster schools demonstrating mean scale scores below expectations in reading and math included Griffith ES, Martinez ES, and McWilliams ES with Fyfe ES performing below expectations in reading only and McCaw ES performing below expectations in math only. The remaining schools identified as enrolling larger than average IEP populations had mean reading or mean math scale scores that were similar.

Conclusions

The No Child Left Behind Act was introduced to address the achievement gap that exists between certain ethnic and special needs populations. In its current form, NCLB requires standardized methodologies to annually evaluate the performance of
schools regardless of student profile or circumstances. Included in this evaluation is the performance of all major subgroups, which include American Indian, Asian, Hispanic, African American, White, IEP, LEP and FRL. While designed to narrow the achievement gap, with emphasis placed on eventually eliminating the academic disparity between minority and nonminority students, as well as disadvantaged children and their more advantaged peers, no provisions were included to accommodate the achievement gap that currently exists.

This work in no way suggests NCLB has had a negative impact on schools or districts, and in fact supports the program. In a September 25, 2007 letter from the U.S. Department of Education, Secretary of Education Margaret Spelling commends educators on the significant progress in reading and math since the implementation of NCLB. She states “Student achievement is on the rise. Any efforts to weaken accountability would fly in the face of rising achievement”. She further points out that 48 states and the District of Columbia have either improved or held steady in reading and math achievement since the implementation of NCLB (Spelling, 2007).

While not an attack on NCLB, this work nonetheless identifies weaknesses that must be addressed. Anyone with an elementary knowledge of statistics can accept the overall gains suggested by Secretary Spelling. Unfortunately, the U. S. Department of Education tends to focus on aggregated groups of students and schools, with little attention being placed on individual schools enrolling large minority or special-needs populations.

Schools were evaluated under NCLB guidelines for the first time at the close of the 2002-2003 school year, with the fifth year of a school being identification as in need
of improvement a trigger for preparation for restructuring. Currently, Nevada has not
defined specific guidelines for restructuring, with options including reopening as a
charter school, reconstitution (replacing all or most of staff), contracting with a
management company such as Edison Schools, turning over operations to the State
Department of Education, or other methods consistent with NCLB (Northwest Regional
Educational Laboratory, 2005). Regardless of the procedure selected, the language of
NCLB requires states to enforce at least one of the restructuring options during the sixth
consecutive year of not making AYP.

The 2007-2008 school year marks the sixth year of evaluation under NCLB,
meaning schools that have consistently failed to achieve AYP will be entering needs
improvement year 5 at the close of the school year. Regardless of the option Nevada
Department of Education selects for restructuring, controversy will most certainly follow.
While not specifically investigated, many of the schools with the potential to enter
restructuring at the close of the 2007-2008 school year may in fact be identified in this
study as performing above expectations relative to populations served. Had alternative
methods been considered throughout the program, schools may have been identified as
not making AYP under standardized procedures, but nonetheless narrowed the
achievement gap, which could have possibly delayed restructuring procedures.

As this study has demonstrated, the use of mean reading and mean math scale
scores instead of percent proficient can provide data that better defines the achievement
gap, which would also assist in quantifying the academic performance of schools
enrolling similar populations. Comparing the current rate of proficiency for schools at
the lower end of the achievement gap to the universal targeted rate of proficiency may
result in schools appearing to perform below standards whereas comparing mean reading or mean math scale scores relative to schools with similar populations may provide an opposite outcome. It is important to consider schools with large, homogeneous minority populations also enroll, and therefore commonly report, the performance of special populations such as IEP, LEP, and FRL, which are simply small samples from the larger minority groups. While the probability of not making AYP increases with each additional reported subgroup, subgroup reporting rules place schools enrolling large diverse populations at a disadvantage.

NCLB is scheduled for reauthorization during the 2007-2008 school year. In a report released by the Government Printing Office titled “Building on Results: A Blueprint for Strengthening the No Child Left Behind Act (GPO/LPS79571, 2007)“, Secretary Spelling offers suggestions to consider during reauthorization. One such suggestion is the use of growth models, which would give credit to students and schools that demonstrate a narrowing of the achievement gap. This study has established a method that would create baseline values and possibly monitor progress. The single drawback would be methods that are not understandable and interpretable to the stakeholders. The current use of percent proficient is easily understood whereas introducing a growth component such as the one used in this study adds a level of complexity that could make interpretation of results difficult.

Also included in Secretary Spelling’s reauthorization comments are provisions for addressing the achievement of non-English speaking students and students with IEP’s. It has been suggested schools will receive recognition for making significant progress in moving students toward English proficiency with IEP students possibly given the option
of testing using modified achievement standards. Both suggestions directly address deficiencies identified in this work and would possibly rectify issues encountered with non-English speaking and special education students.

While acknowledging an achievement gap exists, NCLB provides no provisions to accommodate schools performing at the lower end of the achievement scale, making it difficult if not impossible for low performing schools to catch-up within the allocated time frame of on or before the 2013-2014 school year. It is important to be cognizant of across-subgroup performance, and manage an already existing achievement gap, while at the same time recognizing achievement within select subgroups. Valuable information is lost when recognition is not applied to schools operating within a few city blocks of one another, enrolling similar groups of students, yet performing at opposite ends of the academic spectrum.

**Recommendations for Further Study**

1. This study has identified groups of schools as similar in all respects with the exception of achievement. Schools located in narrow geographic areas, enrolling similar groups of students, should not demonstrate statistically significant differences in achievement. Any situation where the within-cluster achievement differs significantly suggests something other than student demographics may account for the disparity. A qualitative research study could be designed that would investigate the quality of relationships, unique activities, distinctive situations, and academic programs at schools demonstrating significant differences in mean reading or mean math scale score.
A recommendation for further study could begin with identification of the phenomenon to be studied, in this case the disparity in reading or math scale scores, followed with identifying the subgroup cluster and schools to be studied. The study should progress with formulation of a preliminary hypothesis, collect data, analyze the collected data, and draw conclusions based upon findings. Considering the schools under study demonstrate similar demographic profiles, drawn from narrow geographic areas, the assumption can be made all confounding variables outside the school environment will remain constant. Therefore, causation should be confined to factors related to the school environment such as school culture, academic programs adopted, administrative practices, staff experience, or leadership style. Of course, this is not an exhaustive list as there could be a variety of factors interfering with student academic success or failure.

2. Each year schools are assigned a designation based upon performance on state mandated testing. The designation could be adequate if the school was proficient in all areas, watch if it is their first year of not meeting standards as required through NCLB, or in need of improvement year n, with n dependent upon how many consecutive years the school has not achieved standards. The designation is assigned based upon percent proficient, using standard adopted methodologies, with no consideration for population served. This study performed a similar evaluation, using the mean reading and mean math scale scores from the same data set, with recognition for achieving above or below expectations based upon population served. It is quite possible the alternative approach in this study had an influence on school designations.
As we enter the sixth year of NCLB, certain schools will be in a position to face restructuring. A recommendation for further study might include comparing and contrasting designations based upon the two methods outlined in this study. Using student data from the 2002-2003, 2003-2004, and 2004-2005 school years, cluster membership could be confirmed or denied based upon demographic changes and, using mean reading and mean math scale scores, inter-cluster school progress could be tracked. Schools that consistently perform above expectations, relative to respective cluster, should also have a favorable designation, as determined through NCLB. Conversely, schools that consistently perform below expectations, relative to respective cluster, should have an unfavorable designation, as determined through NCLB. This study would confirm or deny the methodologies currently used to evaluate schools under NCLB reflect comparable results when evaluating schools based upon performance relative to schools with similar demographic profiles.

3. In this study the Hispanic subgroup was by far the largest minority group, representing 38.67% of the assessed population, with non-English speaking students representing 27.80% of the assessed population. As was demonstrated earlier, the correlation between Hispanic and LEP was strong enough to combine subgroups, with Hispanic representative of Hispanic and LEP. Testing non-English speaking students with English-only assessments introduces construct irrelevant components that reflects in part current levels of English proficiency instead of reading or math abilities (American Educational Research Association, 2004). The practice of testing non-English speaking students with English only assessments is further reflected in standard 9.1 of the Standards for Educational and Psychological Testing (American
Educational Research Association, 2004) which states “Testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences,” meaning any inferences introduced concerning a non-English speaking students abilities are inappropriate when using English-only assessments.

Considering the combination of a large Hispanic population and the correlation between Hispanic and LEP, a recommendation for further study could include a reliability and validity test of the Hispanic subgroup. The suggestion includes Hispanic only as an argument can already be made concerning the lack of reliability and validity when testing non-English speaking students with an English only assessment. Further testing to verify that fact seems pointless.

Test reliability could be measured using one of two methods. The test-retest procedure simply requires administering two similar but different tests and recording the correlation between them. A strong enough correlation would signify reliability (Vogt, 1999). The second method would be the somewhat more complicated test of internal consistency, which involves testing the correlation between groups of equivalent questions and computing the correlation using Cronbach’s Alpha (Vogt, 1999). Validity refers to the degree the measurement instrument is accurately measuring what it was designed to measure. More specifically, construct validity refers to how well an assessment can be interpreted as a measure of an attribute or quality for which it is operationally defined (Vogt, 1999). English-only reading and math tests are designed to test the reading and math skills of English-speaking students and will not test the reading and math abilities of students that non-English
or limited English speaking. While tests to measure construct validity can be quite complicated, a simple suggestion might be a correlation test between the reading and math tCRT and the Language Assessment Scales (LAS), an assessment administered annually to all CCSD non-English speaking students. A positive correlation between performance on the reading or math CRT to the LAS would provide the operational measure necessary to verify whether construct validity does or does not exist.
REFERENCES


inequalities in school achievement. *Journal of Marriage and the Family, 60*(3),
715-723.

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2002). *The analysis and
interpretation of multivariate data for social scientists.* Boca Raton, Fl: Chapman
& Hall.

Time, Family Composition, and the Acquisition of Social Capital. *Journal of
Marriage and the Family, 59*(2), 332-344.

Chubb, J., Linn, R., Haycock, K., & Wiener, R. (2005, Spring). Do we need to repair the
monument? Debating the future of No Child Left Behind. *Education Week, 5*(2),
8-19.

Coladarci, T. (2003). *Gallup goes to school: the importance of confidence intervals for
evaluating “adequate yearly progress” in small schools.* Washington, D.C.: Rural
School and Community Trust.

analysis for performance assessment of student’s achievement or school

the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99, Fall
National Center for Educational Statistics.

181


186


VITA

Graduate College
University of Nevada, Las Vegas

Jeffrey Neal Halsell

Local Address:
603 Ancient Mayan Drive
Henderson, Nevada 89015

Degrees:
Bachelor of Science, 1994
University of Nevada, Las Vegas

Master of Arts, 1999
University of Nevada, Las Vegas

Dissertation Title: Using Cluster Analysis to Evaluate the Academic Performance of Demographic Homogeneous Subsets

Dissertation Examination Committee:
Co-Chairperson, Dr. Robert S. McCord, Ed. D.
Co-Chairperson, Dr. Chad L. Cross, Ph. D.
Committee Member, Dr. James R. Crawford, Ph. D.
Graduate Faculty Representative, Dr. Lawrence R. Walker, Ph. D.