GRADES AND DATA DRIVEN DECISION MAKING: ISSUES OF VARIANCE AND STUDENT PATTERNS

By

Alex Jon Bowers

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Educational Administration

2007

ABSTRACT

GRADES AND DATA DRIVEN DECISION MAKING: ISSUES OF VARIANCE AND STUDENT PATTERNS

By

Alex Jon Bowers

This study addresses the question: *To what extent are teacher assigned subject-specific grades useful for data driven decision making in schools*? Recently, schools have been urged to bring teachers and school leaders together around student-level data in an effort to increase dialogue, collaboration and professional communities to improve educational practice through data driven decision making. However, schools are inundated with data. While much attention has been paid to the use and reporting of standardized test scores in policy, school and district-level data driven decision making, much of the industry of schools is devoted to the generation and reporting of grades. Historically, little attention has been paid to student grades and grade patterns and their use in predicting student performance, standardized assessment scores and on-time graduation. This study analyzed the entire K-12 subject-specific grading and assessment histories of two cohorts in two separate school districts through correlations and a novel application of cluster analysis. Results suggest that longitudinal K-12 grading histories are useful. Grades and standardized assessments appear to be converging over time for one of the two school districts studied, suggesting that for one of the districts but not the other, current accountability policies and state curriculum frameworks may be pushing into classrooms and modifying teacher's daily practice, as measured through an increasing correlation of grades and standardized assessments. Moreover, using cluster analysis, K-12 subject specific grading patterns appear to show that early elementary

school grade patterns predict future student grade patterns as well as qualitative student outcomes, such as on-time graduation. The findings of this study also suggest that K-12 subject specific grade patterning using cluster analysis is an advance over past methods of predicting students at-risk of dropping out of school. Additionally, the evidence supports a finding that grades may be an assessment of both academic knowledge and a student's ability to negotiate the social processes of school.

Copyright by

ALEX JON BOWERS

2007

Bowers, A.J. (2007)

Bowers, A.J. (2007)

ACKNOWLEDGEMENTS

The journey through the Ph.D. process at Michigan State has truly been an excited and fascinating time for me. Many people have helped me along this path, and without each of them, my time and the quality of my work would not have been as successful.

I would like to thank my dissertation committee, Susan Printy, Philip Cusick, BetsAnn Smith and Kimberly Maier for their time, their thoughtful comments, their patience and their willingness to let me pursue the path less traveled. Specifically, I thank my advisor Susan Printy, for her time, her careful reflection on my ideas and reading of my writing, and her never ending understanding and willingness to let me pursue a research agenda that I believed in. I thank Philip Cusick for his constant faith in me and my career and his willingness to take a chance on a past molecular biologist who wanted to get a Ph.D. in educational leadership and do good work in schools. I thank BetsAnn Smith for her thoughtful feedback, and on always being there to provide constructive guidance. I also thank Kimberly Maier for her expert assistance and critical feedback.

Additionally, I would like to thank those who have helped me throughout my time at MSU. I thank Gary Sykes for his moral, intellectual and financial support of me and this project. Learning about educational research from Gary, both in classes and in collaboration on research projects, has been a highlight for me. I also would like to thank the Educational Administration Department at MSU for supporting my research, giving me the opportunity to teach while at MSU, and making it possible for me to pursue my research.

I also thank my family. I thank my parents, John and Ann Bowers for their unrelenting belief and support of me. I also thank William and Patricia Nurnberger for their insight and encouragement. And, I wish to thank both of my grandmothers, Beatrice Peterson and Thelma Bowers, for their life-long support and love of learning. They both, unfortunately, were not able to see me through to the end of this journey.

Most of all, I thank my wife, Amy Nurnberger. Your support, encouragement, passion, understanding, and love sustain and energize me. You make hard work fun, you listen to my ideas, you question my conclusions, and you provide support when I need it the most. You bring joy to my life and make every day interesting and fun.

TABLE OF CONTENTS

Bowers, A.J. (2007)

LIST OF TABLES

LIST OF FIGURES

KEY TO SYMBOLS OR ABBREVIATIONS

GPA	Grade Point Average
HSGPA	High School Grade Point Average
K	Kindergarten
NOTG	Not On Time Graduation
10 S-2	10^{th} Grade Semester 2
SES	Socioeconomic Status
SSF	Success at School Factor

Bowers, A.J. (2007)

CHAPTER I: INTRODUCTION

The use of data to inform the decision making of school leaders and teachers in K-12 American schools continues to be a topic emphasized not only by organizational researchers who see data driven decision making as a means of instructional improvement (Bernhardt, 2004; Coburn & Talbert, 2006; Halverson *et al.*, 2005; Kerr *et al.*, 2006; Raudenbush, 2005; Streifer, 2004; Thorn, 2002; Wayman & Stringfield, 2006a; V. M. Young, 2006), but also according to law and policy as stricter mandates have been passed requiring data reporting and evidence based practice in schools (Earle & Fullan, 2003). Schools are inundated with data, including grades, attendance, discipline records, and standardized test scores (Creighton, 2001a). While much attention has been paid to using standardized test scores for data driven decision making (Bernhardt, 2004; Streifer, 2004), much of the industry of schools is devoted to grades, creating a dualistic system: one based on standardized testing and decision making that reports to policymakers and the government, the other based on grades that reports to students, parents and the community (Farr, 2000). Thus, the question for this study is*, can grades be used for data driven decision making?*

Historically, grades have been criticized for being subjective and unreliable measures of student achievement (Cross & Frary, 1999; Kirschenbaum *et al.*, 1971). While standardized assessments have undergone a "virtual revolution" over the past thirty years in reliability and validity of measuring student academic achievement (Cizek, 2000), no such revolution has occurred in the arena of grades (Cizek *et al.*, 1995-1996; Trumbull, 2000b). If grades are subjective and unreliable, how do they fit into a discussion of data-driven decision making for school improvement? One approach

Bowers, A.J. (2007)

identified in the literature shifts emphasis away from the criticism of the subjectivity of grades to a discussion of ways of making grades more valid through triangulation and cross-referencing grading data with numerous other data sources in schools (Bernhardt, 2004), and aligning grading with state curriculum standards (Farr, 2000). Thus, one of the hypotheses of this study is that while grades may have been subjective and unreliable assessments in the past, it may be that currently as schools are pressured to align assessments with state mandated criterion and curriculum, the two systems of grades and standard assessments are converging into one, increasing the correlation between the two assessment systems.

One theory proposed for past grade subjectivity has focused on the influence of teacher and student perceptions on student grades. It is hypothesized that students who receive high grades in early elementary school continue to receive high grades throughout their schooling career due to the positive motivation of high grades and teacher and student perception of student ability based on past student performance (Hargis, 1990), termed here the "Hargris hypothesis." Moreover, it is hypothesized that students who are given low grades early on are locked into a cycle of low grading. However, the question of how student's grades pattern over time has not been empirically tested to date. If past grading patterns are predictive of future student grade patterns, this would allow school leaders to predict future student grade performance outcomes (such as in high school) in elementary school in specific subjects, and thus design instructional interventions for individual students in specific subjects before they become locked into a cycle of low grading patterns with a higher probability of dropping out of school.

Hence, the research questions for this study are: 1) To what extent has the correlation between grades and standardized assessments changed over time? 2) To what extent does the hypothesis that past student grade patterns predict future student grade patterns hold true? 3) To what extent is grade patterning useful in predicting student outcomes such as graduation or dropping out? To what extent do these predictive patterns aid in identifying avenues for early intervention by instructional leaders and teachers?

This study outlines two domains for research within the broader issue of using grades as data for decision making by educational leaders. First, to explore the possibility that grades and standardized assessments may be converging, subject specific grades and standardized state assessment scores were correlated for the 1994 and 2006 graduating cohorts from two separate K-12 school districts. The evidence suggests that subject specific grades and standardized assessments may be converging for one of the two districts. This may be an indication that assessment policies may have affected one of these two school districts, but not the other.

Second, a novel application of hierarchical cluster analysis is used to explore whether early student grade patterns are predictive of future student grade patterns and if overall student grade patterns are predictive of qualitative student outcomes, such as on-time graduation. The data suggests that generally, early student grade patterns are predictive of future student grade patterns. The application of cluster analysis to longitudinal subject-specific K-12 student grade data allows for the identification of specific timepoints in early elementary school for multiple clusters of students that may be important in the decision of where to apply the limited resources of a school district for data driven decision making by educational leaders. Additionally, cluster analysis of

subject-specific grades appears to be an advance over past methods of prediction of students at-risk of not graduating on time, not only using K-12 grade data, but interestingly also K-8, K-6, and even K-1. Furthermore, the evidence suggests that grades may be an assessment of both academic knowledge and a student's ability to negotiate the social processes of school.

Through the analysis of K-12 subject-specific grades and standardized assessments, the contention of this study is that teacher assigned subject specific grades are important and useful as data for data driven decision making by educational leaders. Furthermore, as data that schools already collect on students, grades may predict future student outcomes, providing grade-level and subject-specific intervention points for school and district-level data driven decision making.

CHAPTER II: LITERATURE REVIEW

Using Data for Instructional Improvement

Over the forty years since the publication of the Coleman report (Coleman *et al.*, 1966) and as the demands of the accountability movement have gradually increased, schools and school districts have increasingly come under pressure to improve performance through the use of data analysis (Fullan, 2000) to the point where data analysis in schools has become unavoidable (Earle & Fullan, 2003). Currently, much of the literature urges school leaders and decision makers to use data-driven decision making to help inform their practice and help them make sound decisions based on what the data in their schools tells them (Bernhardt, 2004; Halverson *et al.*, 2005; Streifer, 2004; Wayman, 2005; Wayman & Stringfield, 2006a).

It has been argued by Elmore that the relatively recent increase in accountability and performance pressures on the educational system from external agencies is due to a switch from what he terms the "attainment culture" to the "performance culture" (Elmore, 2002, 2003). In the attainment culture, schools were judged by how well children who were deemed worthy of an education were moved through the system. However, beginning with standardized tests in the 1960s, and continuing to this day, vast inequities were realized within the system, revealing the large differences in test scores and knowledge acquisition not only between children of high SES families and low SES families but also between different ethnic groups. With this realization by businesses, government, and policy makers, along with the general rise of performance-based methods and organizations in the general society, schools have been pressured to change to a performance culture. What this means is that political leaders link funding to

progress by the educational system in an attempt to hold the educational system accountable for the learning of the historically lower performing SES and ethnic groups revealed by standardized testing. Moreover, this performance culture has risen concurrent with, and is likely linked with, school and district attempts to incorporate the tenets of quality management, including continuous improvement, customer focus, systems thinking, process evaluation and data-driven decision making (Detert et al., 2000). This performance and quality management culture has shifted the processes of schools from the education of a *selection of students with low accountability* to the public to the education of *all students with high accountability* to the public, creating a need for schools and districts to examine their data closely to determine what decisions to make about what works in schools (Fullan, 2001; Raudenbush, 2005).

The Challenges of Using Data in the School Context to Make Decisions

With the advent of the performance culture and quality management, schools have begun to focus on collecting, discussing and using data to inform decision making processes (Bernhardt, 2004; Coburn & Talbert, 2006; Halverson et al., 2005; Kerr *et al.*, 2006; Streifer, 2004; Thorn, 2002; Wayman & Stringfield, 2006a; V. M. Young, 2006). While in the past, decisions by school leaders oftentimes were based on intuition, fads, rules of thumb, or past experience (Bernhardt, 2004; Creighton, 2001a; Earle & Fullan, 2003), exemplary schools and school districts have been shown to use data effectively to improve instruction (Edmonds, 1979; Elmore & Burney, 1999; Hightower & McLaughlin, 2005; Kerr *et al.*, 2006; Massell & Goertz, 2002; Schmoker, 1999). For these effective schools, data use, through monitoring of student academic progress and intervention for individual students, was one of five factors that also included a focus on

basic skills, high expectations for all students, strong instructional leadership and an orderly environment (Murphy & Hallinger, 2001; Teddlie, 1994). While multiple examples of high performing low income schools exist, "what works" (Raudenbush, 2005) remains a question for leaders in schools and districts.

For many educational leaders working from a quality management perspective, the question of what works motivates them to examine the effects of the organization on the students, and determine the causes of those effects (Supovitz, 2002). It has been well argued that the gold standard for determining causality is randomized controlled trials (Raudenbush, 2005). Only through random assignment of treatment and controls is a researcher able to say with certainty if a specific intervention caused an outcome. However, for most schools, it is prohibitively expensive to conduct such trials (Raudenbush, 2005). While some authors have argued that school districts randomly assigning scarce resources and then tracking the outcomes over long periods of time is not only possible, but has succeeded in the past (such as in the Perry preschool study) (Rothstein, 2004), for the vast majority of schools, random controlled experiments are beyond the scope of their expertise and funding (Streifer, 2004). Thus, school leaders rely on specific statistical techniques to aid them in using data effectively.

Most often, the next best technique is multiple regression statistical analysis (Streifer, 2002). Using multiple regression, an evaluator is able to take the vast variety of data generated by students and use it to predict future student outcomes on specific variables, such as state test scores (Streifer, 2004). However, using this technique in schools violates many of the assumptions of multiple regression, including large enough sample sizes, the independence of cases, the independence of variables (multicolinearity),

normality of the data, the independence of variance explained from the variance remaining, and the homogeneity of variance across cases (Cohen *et al.*, 2003; Howell, 2002; Rencher, 2002). Though additional statistical methods such as data mining algorithms (Streifer, 2004, 2005) or hierarchical linear regression techniques (such as HLM) (Raudenbush & Bryk, 2002), address some of these issues with multiple regression (namely multicolinearity and the dependence of cases), the other issues remain, leaving multiple regression as a poor statistical procedure for use by school leaders. Furthermore, inferential statistical techniques such as multiple regression are designed to estimate the mean for the population from which a sample is taken. If one already possesses all of the data for a selected population (such as a school district), there is no need to estimate the population means since one can calculate them directly. Many school leaders who are looking to determine what works in their schools do not wish to generalize their population of students to the greater population averages, which is the purpose of multiple regression. Rather, they wish to know what is working and is not working for their specific students for the very near future, measured in near-term time-frames (Creighton, 2001a).

Leveraging data to make decisions at the school level is a complicated process (Wayman, 2005). It has been argued that educational leaders should forgo the more difficult and complex issues around higher level statistics and concentrate rather on collecting data from multiple sources, including test scores, grades, demographic data, school processes, community and organization perceptions. They should use descriptive statistics to better understand what the data says for their specific situation, making more informed decisions based on those descriptive reports that create an overall picture of

what is occurring in schools (Bernhardt, 2004; Halverson *et al.*, 2005; Kerr et al., 2006). While the literature points to the types of data to be used, and different ways to analyze that data, it is school leaders who must use that data in decision making processes.

While some schools have led successful improvement efforts, instructional improvement across the system is acknowledged as spotty and in need of much more improvement, especially for children of urban, low SES and ethnic minority families (Elmore, 2002). It has been argued that since the 1970's we have had all of the data needed to improve schooling for not only these subgroups of children, but all children (Edmonds, 1979; Marshall, 1997). Schools, however, are awash in data, generating standardized test scores, achievement scores, grades, attendance, discipline reports and portfolios on each student. Such data can result in a disorganized and incoherent database (Brunner *et al.*, 2005; Cizek, 2000; Earle & Katz, 2003; Salpeter, 2004; Streifer, 2004) presented in dense and inaccessible reports to school leaders and teachers (Wayman *et al.*, 2004) who on average have a rudimentary training in statistics (Creighton, 2001b; Earle & Fullan, 2003; Secada, 2001). For educational leaders in the current era however, linking instructional improvement to a critical analysis of data is now unavoidable (Earle & Fullan, 2003). While many school leaders currently focus on standardized test scores, data is being collected daily on every student in multiple ways (Bernhardt, 2004; Creighton, 2001b). Much of this data collection of schools is centered on grades.

Grades, Grading and Marks

Historically, the vast industry of data collection in schools and school districts has centered on grades. Since its inception, the American public school system has had a focus on grades and grading (Quann, 1983) with the purpose of providing feedback to

administrators and potential employers for predicting student's future performance from current grades, guiding students to areas of aptitude, providing student performance information to parents and administrators, and motivating students to do well and be well disciplined (Evans, 1976; Trumbull, 2000b). For students, working to achieve a high grade, compete against fellow students, or game the system takes up a large percentage of their time. For teachers, designing and proctoring assessments, grading the assessments, and negotiating with students over their grades requires substantial amounts of time both inside and outside of the school day (Hargis, 1990; Kirschenbaum *et al.*, 1971). These unending demands on time concerning grades and grading for both teachers and students are in addition to the relatively recent introduction of standardized testing. With the advent of state and federally mandated testing, schools and school districts are increasingly devoting more and more time to preparing for and administering these state standardized tests (Militello, 2004; Salpeter, 2004). The work surrounding grades and grading however, continues unabated. As a result, in American K-12 education we have a dualistic assessment system (Farr, 2000), one based on psychometrically standardized tests, and one based on the subjective industry of acquiring and awarding grades.

In the past, the practice of standardized testing was criticized for how they were used and for the validity of the tests (Goslin, 1968). More recently, standardized testing has undergone a "virtual revolution" with an increase in test validity and reliability (Cizek, 2000). Unfortunately, no such revolution has occurred in the arena of grades and grading (Cizek, 2000; Cizek *et al.*, 1995-1996; Trumbull, 2000b). For those who have examined grades and grading practices, grades and marks have been reported to be highly variable and subjective, failing to adequately perform the stated purpose of providing

feedback, prediction, guidance, information and motivation for students and their parents (Hargis, 1990; Kirschenbaum et al., 1971; S. Simon, 1976).

A study of four elementary schools in California demonstrated the subjectivity and the role of teacher perception on student achievement. All students were given a test that teachers had been told would predict the IQ gains of the child over the next year. About ten students were selected at random from every classroom and assigned a high score. The teachers were then told only about the scores, not that the children were randomly assigned. These children in each class were then used as the experimental group and the remaining children as controls. At the end of the year, children in the experimental group from kindergarten, first and second grades made significant gains on IQ tests and achievement measures over the controls, and teachers rated the children as more cooperative, more socially adjusted and more well behaved (Rosenthal & Jacobsen, 1969). While dubious ethically, and followed by multiple publications questioning the veracity of the claims of the study (Elashoff & Snow, 1971), the basic assertion that in early grades teacher perception may influence student achievement has been supported (Raudenbush, 1984; Spitz, 1999). Thus, student outcomes may be dependent on teacher perceptions.

Other studies have examined the practice of grading and how teachers construct grades by incorporating many different factors. These practices have been termed "hodgepodge" grading practices (Brookhart, 1991; Cross & Frary, 1999) with little reliability; it is basically random within schools, differentially incorporating academic achievement as well as effort, improvement and behavior into assigned grades (Cross & Frary, 1999; Frary *et al.*, 1993). Identified by Talcott Parsons over 45 years ago, grades

have been recognized indicators of academic, interpersonal and social factors (Parsons, 1959). Recent studies have shown that teachers independently incorporate into grades such personalized measures as classroom participation, attendance, behavior and conduct, completion of homework, achievement on homework, student ability, student growth and improvement, effort, and achievement on classroom assessments (Cross & Frary, 1999). One could argue, then, that what a grade represents is different for different teachers and different students within the same school building. With such a system of grades, what a single letter grade represents is unknown.

In addition to their dependence on perception and to their hodgepodge grading nature, grades have long been criticized as essentially subjective. The seminal studies of Starch and Elliot demonstrated the subjectivity in teacher graded English, geometry and history exams (Starch & Elliot, 1912, 1913a, 1913b). In their first study, the researchers took two English exam questions and answers, and sent the sets to 200 schools requesting that the head of the English department grade the answers on a 100 point scale. Of the approximately 150 exams returned, the range in scoring was about 39 points for both, meaning that while some teachers gave a high "A", others gave a "D" to the same answers. Once this study was published, an outcry arose that English was a subjective subject, so a large range on an English exam could be expected (S. B. Simon & Bellanca, 1976). Subsequently, Starch and Elliot attempted the same study with geometry. Of 138 geometry exams graded and returned, the range in scores on a one hundred point scale was 45 points, even greater than the English exam (Starch & Elliot, 1913a). A replication for a history exam also gave a range of 40 points (Starch & Elliot, 1913b). Evaluation of the comments given by the graders showed that teachers scored the exams very

differently, marking for a combination of penmanship, neatness, spelling, showing all work, and the right answer, no matter the subject area. The scores returned represented a normal curve; the same result one would expect sampling a random population. These results indicated that exam scoring is subjective, and variation among teachers is random. Although much has been done to increase the objectivity, reliability and validity of standardized tests, little to no progress in the area of grading objectivity and reliability has been observed in the 100 years since these problems were first described (Cizek, 2000; Kirschenbaum et al., 1971).

The reasons behind this subjectivity have been sparsely addressed in the literature. For those who have studied it, this persistent problem of grades has been attributed to teacher isolation (Elmore, 2002), little to no training or preparation in testing and measurement for teachers (Carr, 2000), and a lack of dialogue and communication between teachers and district administration about grading standards and alignment (Cross & Frary, 1999; Kirschenbaum et al., 1971). Furthermore, bias in assessment and scoring is thought to be widespread (Trumbull, 2000a). Combined with the subjectivity of grades discussed above, these issues add to the list of problems with grades.

Three current theories provide insight into issues related to grade subjectivity. First, it has been noted that the practice of grouping children in grades based on their age, with an arbitrary cutoff date for yearly enrollment, generates classrooms that are assumed to have low variance of pre-knowledge among the children. However, in fact, the children have been shown to differ by as much as three grade equivalent years of knowledge upon entrance to first grade, a gap that continues to increase in subsequent years (Hargis, 1990). To cope with the vast variance contained within a classroom,

teachers provide one level of instruction, directed to the median knowledge-base and ability of the class (Hargis, 1990). Second, as a counter to the known subjectivity of teacher assigned grades and to increase the statistical calculation of test reliability, psychometricians have instructed teachers to increase classroom variance by placing extremely difficult questions on tests. This practice increases the ranking capability of the tests (Carr, 2000; Cross & Frary, 1999), and also ensures that some portion of students will have difficulty or may fail. Third, while empirical data is sparse, it has been hypothesized that children who receive high grades continue to receive high grades throughout the schooling process, and children who receive low grades continue to receive low grades due to the positive motivation of high grades, the absence of motivation of low grades, teacher perceptions of student ability based on past grades, and the ability tracking assigned to grades by the organization (Evans, 1976; Hargis, 1990; Kirschenbaum et al., 1971). Combined, these three theories indicate that the traditional grading system ensures that a certain percentage of students will fail, as children are graded, ranked and tracked through the system (Hargis, 1990). This is especially troubling given the above discussion of the subjectivity and "hodge-podge" nature of teacher assigned grades.

Using Grades for Data-Driven Decision Making

The question that underscores this study is: *If grades are subjective, invalid and unreliable, how do they fit into a discussion of data-driven decision making for school improvement?* While it has been argued that due to these problems with grading, grades could be eliminated from schools and students could be judged only on standard

assessments (Kohn, 1994), it is understood that grades are an integral part of the function, structure, and community perception of schools and are thus, here to stay (Hargis, 1990; Kirschenbaum et al., 1971; Trumbull, 2000b). Thus, the emphasis has shifted to a discussion of ways of making grading more "instructionally valid" (Newmann, 1991), triangulating and cross-referencing grading data with the numerous other data sources in schools (Bernhardt, 2004), and aligning grading with state curriculum standards and standardized tests (Carr & Farr, 2000; Farr, 2000; Waters, 2000). For teachers, however, grades have "face validity"; teachers are often more willing to accept grades over other assessments such as standardized tests because they assigned those grades based on their own assessments (*Ncrel guide to using data*, 2004; Mehrens & Lehmann, 1991).

For schools and school districts, analyzing grading data for decision making is vitally important. Despite all of the known issues with grades and grading subjectivity addressed above, grades are used to make decisions that have direct impact on both students and schools. Grades are used to make decisions for special needs testing, to assign special education services, and to admit or channel students into specific curriculum tracks (Hargis, 1990; Langdon & Trumbull, 2000). For schools, these decisions impact not only finances, especially with special education decisions, but also the long-term success of students, including dropout, graduation and college admittance levels. As a result, it is crucially important that schools and districts examine grade data when making decisions that will impact the long-term success of students. The question of exactly how to examine that data in ways to help schools address these issues remains.

CHAPTER III: THEORETICAL FRAMEWORK

While many issues concerning grades and grading need to be addressed, this study focuses on two specific issues related to grades and their potential use in data-driven decision making for instructional improvement in schools. The first issue relates to the hypothesis of grade patterning and the second issue concerns classroom grade variance.

Student Grade Patterning: Identification, Prediction, and Intervention

As referred to above, the supposition has previously been made that due to the subjectivity of grades and the influence of teacher perceptions on grades, students who obtain high grades early on in schooling continue to get high grades throughout their school career and students assigned low grades may become trapped in a cycle of low expectations and grades (Hargis, 1990), termed the "Hargris hypothesis" for this study. It has also been postulated that student motivation, one of the primary goals of grades, only influences students who get high grades (Evans, 1976; Kirschenbaum et al., 1971). The literature on the effects of teacher perception and expectancy on student gains supports this theory of early success, in that if positive teacher perception of a student's ability does influence student gains, then that perception has the most influence in the early grades at the earliest times in the school year (Spitz, 1999). This idea of general early student grade patterns predicting future student grade patterns is shown for a hypothetical dataset of 8 students in Figure 1. This idea of student patterning, the Hargris hypothesis, has been detailed in the literature. Essentially, students who receive high grades in early elementary school are the students who continue to receive high grades throughout their

time in a school district (*Figure 1*, Students 3, 5 & 7), and students who receive low grades early on, may be locked into a cycle of low grading (*Figure 1,* Students 1, 2 & 6). These overall grade patterns have not been empirically demonstrated in the literature to occur over multiple years of schooling.

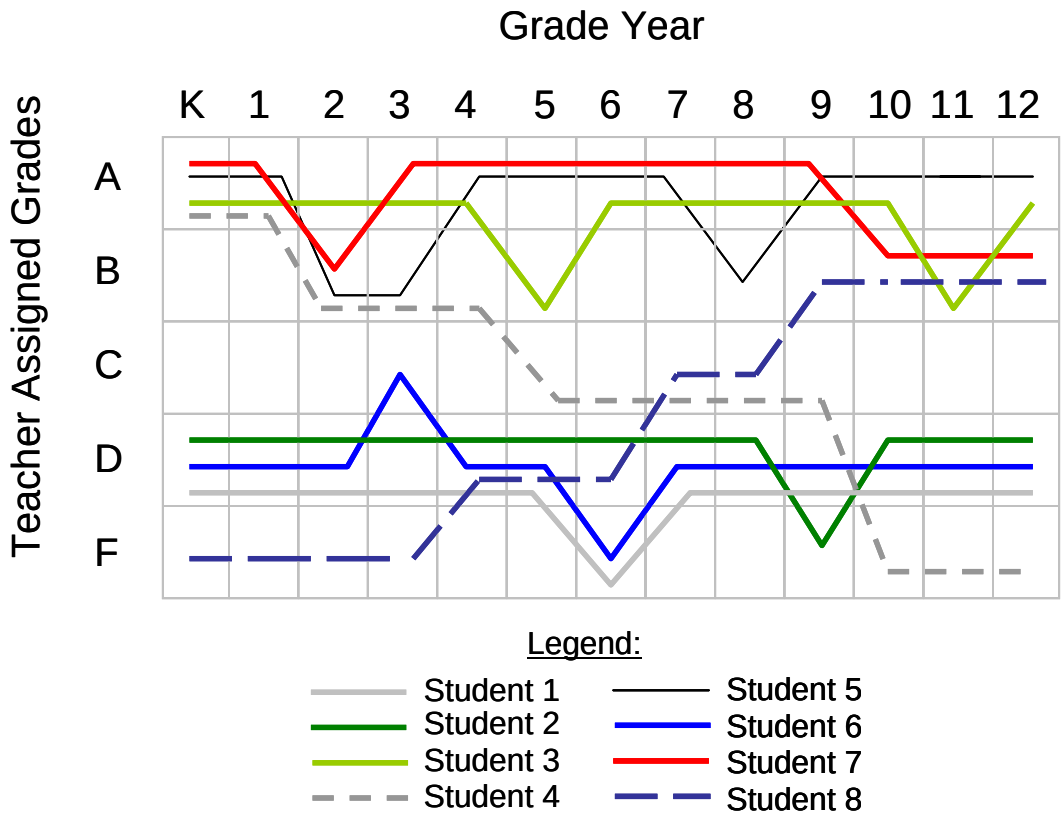


Figure 1: ***Theoretical grading trends for one 13 year cohort in mathematics.*** Although not empirically tested to date, the theory drawn from the literature of the Hargris hypothesis indicates that early on, students 3, 5 and 7 receive high grades and are thus motivated to continue to receive high grades throughout. Students 1, 2 and 6 initially receive low grades and are thus trapped in a cluster of low grading throughout their schooling career. It is unknown if students 4 and 8 exist on a large scale, whether they start high or low and finish in an opposite position, as well as how these different clusters of student patterns are similar and different from each other. More specifically, could student 4 have been recognized in 4th or 5th grade and had an instructional intervention designed to move the student back into the high scoring cluster? **This figure is presented in color.**

If grade patterning in this fashion is occurring, and is more a function of subjective factors rather than actual student achievement (potential or realized), a better understanding of this type of previously unexplored grouping behavior could assist a school district in making more informed decisions about which students are considered underperforming, tracked into special needs assessment, or given access to gifted programs. Furthermore, a better understanding of how student grades pattern with other students within a classroom, cohort, school and district, combined with qualitative data such as district transfer status, gender, retention records, and test taking patterns, could help school leaders pinpoint previously unknown empirically derived subgroups of children who are in need of targeted interventions (*Figure 1*, Student 4). Such data could help inform teachers and administrators of which groups of children are succeeding or failing within the grading system, and what those children's similarities are, in an effort to analyze what works and does not work in a district. Such information would enable a school district to help more children be successful.

A potential statistical tool that could be used to study this type of group patterning is cluster analysis (Lorr, 1983; Rencher, 2002; Romesburg, 1984). In cluster analysis, group patterns can be empirically derived from both grading and standardized achievement test data. Group pattern trends can be used to predict future outcomes, such as using elementary school grades to examine whether or not the Hargris hypothesis is accurate. Another possible use is examining past grade patterns to predict qualitative student outcomes, such as on-time graduation. Since cluster analysis is rarely used in educational research, it shall be explicated at length in the methods.

The Convergence of Grades and Standard Assessments

The second issue addressed in this study is the issue of teacher induced classroom grade variance and the correlation of grades and standard assessments over time. As discussed above, the supposition has been made that teachers are confronted by student populations that have high variability in pre-knowledge and ability, and throughout the course of schooling the variability between students increases. This is attributed in part to teachers using one level of instruction (directed at the middle) and designing assessments that increase variability, each combined with the subjectivity and grouping patterns of grades (Hargris, 1990) discussed above. Even in the case of newer assessment strategies, such as portfolios or formative assessments used in combination with traditional assessments (Airasian, 1994), these issues of teacher subjectivity, perceptions and grade variance, remain. However, with the rise of standard assessments and accountability, it is possible that a currently unexamined change is underway in instruction and teacher grading variance.

As schools and school districts adapt to the introduction of state mandated standardized assessments, they are beginning to realign their curriculum to the state standards under community pressure to perform well as an organization on these assessments. By aligning grade report cards to standardized assessment reports, schools decrease the difference between the two reporting systems (Bisesi *et al.*, 2000; Carr & Farr, 2000). Due to the criterion referenced nature of state standardized tests, teachers must adjust instruction to cover the curricular objectives that the test assesses (Falk, 2002; Popham, 2004). Through alignment of curriculum to the tests, grades and standardized assessments may be converging into one system (Farr, 2000; Linn, 1982,

2000). One hypothesis for this study is that teacher assessment writing, instruction and grading practices for core subjects at all grade levels are changing such that the variance between student grades is decreased, and the usefulness of grades in predicting standardized measures of academic achievement is increased. As teachers personalize and differentiate instruction for students who are perceived to be below the state curriculum criterion and grade students with teacher designed assessments that are aligned with the state standardized assessments, grades may be becoming more aligned with the state standard assessments. If true, this would decrease classroom variance as the low performing students are brought up to the criterion. As a result, grades would be better indicators of student academic knowledge. Of course this hypothesis takes as an assumption that standardized test scores are a valid assessment of student knowledge and academic achievement. A proposal of this study is that grading practices in the current era of accountability, as opposed to grading in the past, are becoming more aligned with standardized assessments as the two systems converge. If this is so, educational leaders could use either grades or standardized assessments to predict each other for the purpose of making decisions at the district, school and student levels.

More specifically, this study investigates the correlation and distribution of grades and standard assessment scores for students within schools and districts at two time points. While not empirically tested, the literature on grades intimates that there may be little correlation between grades and standardized assessments. However, this has not been examined closely in the literature, due in part to the known high subjectivity of grades and the difficulty of obtaining large datasets of student subject-specific and grade-level grading histories.

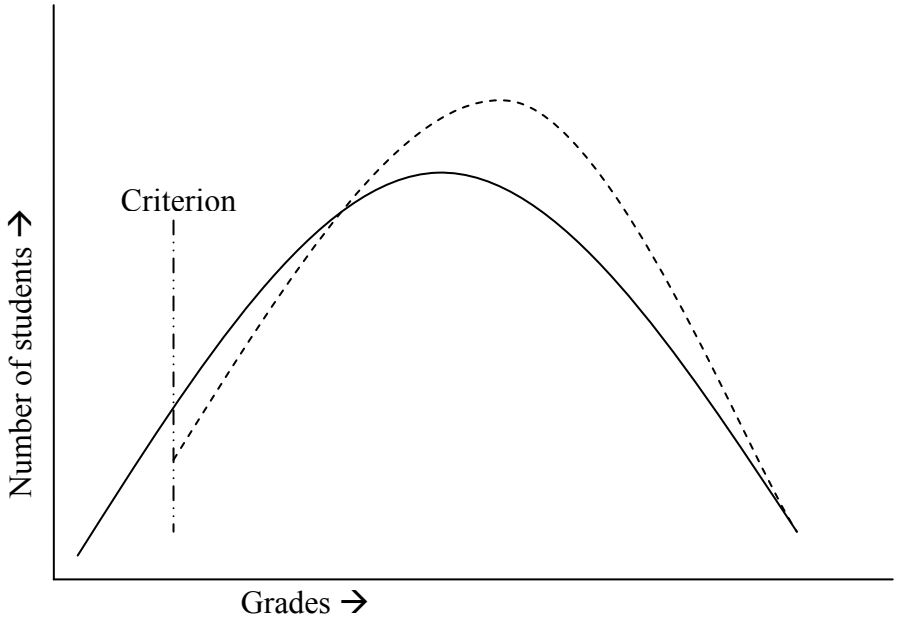


Figure 2: ***Hypothesized change in grade distributions from before and after the implementation of criterion referenced tests.***
Students in the past may have been awarded grades along a normal distribution (solid curve). After the implementation of criterion referenced standardized tests, pacing guides and curriculum alignment, students in classrooms that have obtained 100% passing of the criterion (dashed line) are now hypothesized to have a grade distribution that is skewed somewhat higher (dashed curve).

While also not well studied empirically, the idea that teachers assign grades that distribute students within a normal curve either purposefully or unintentionally, has been hypothesized (Carr, 2000; Cross & Frary, 1999; Hargis, 1990; Kirschenbaum et al., 1971) (*Figure 2*). If teacher grading variance has changed since the introduction of criterion referenced exams and if grades and standardized assessments are becoming more aligned (*Figure 2*), it is a hypothesis of this study that the distribution of grades is beginning to have an increase in positive skew as teachers concentrate instruction on students below

the criterion level, necessarily raising their achievement in all aspects in that classroom, and thus the student's grades (*Figure 2, compare the solid and dashed curve*).

Concurrently, with the implementation of criterion referenced exams and the pressure to align curriculum and classroom practice to state guidelines, grades and standard assessments may be becoming more strongly correlated, in which case, grades and standardized assessments are becoming more predictive of each other. One intent of this study then, is to compare the extent to which grades and standard assessments were correlated at a past date with correlations using current data (*Figure 3*). As shown in Figure 3, currently little is known about the correlation between grades and standard assessments and if that correlation has changed over time. By examining student subject-level grades and state standard assessment scores over time, it can be determined if a change in the correlation of the two assessment systems has taken place over time. However, the cause of the change would still be unknown (*Figure 3*).

If grades have become more correlated with standardized assessments, this would have many implications for schools and districts engaged in data driven decision making. One topical implication would be that districts would have less of a requirement for additional district designed and proctored pre-standardized tests (periodic tests) designed to predict how well a student population will perform on upcoming state mandated tests. Grades alone may sufficiently predict standardized test scores, decreasing the amount of time devoted to pre-standardize assessment preparation, proctoring and evaluation.

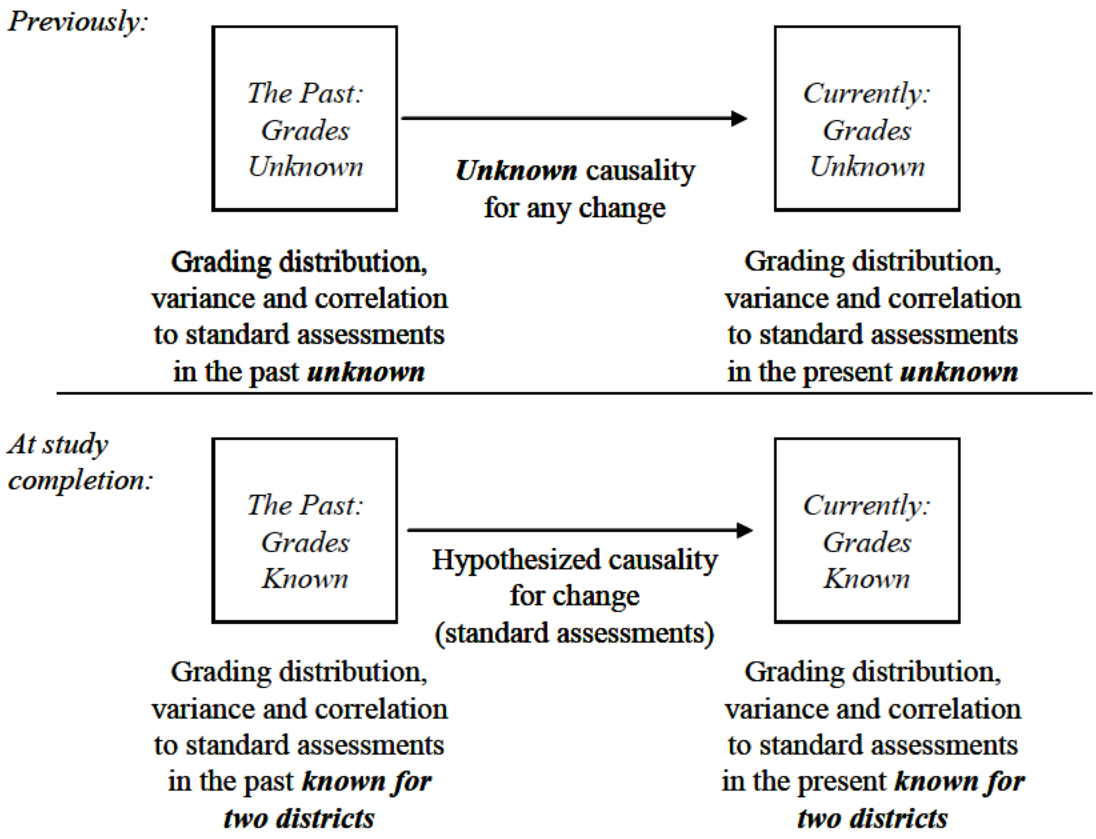


Figure 3: ***Hypothesized scope of the study of the change in grading variance and alignment.***
Previous to the present study, grade distribution, variance and grade correlation to standard assessments for core subjects were unknown both in the past and currently. At completion of the proposed study, grade distribution, variance and correlation to standard assessments will be known for two districts. While causality will not be explored in the study, the hypothesis that standard assessments have led to a change in grade variance is suggested.

Framework Conclusion

Again, the question driving this proposal is: *Can grades be used in data driven decision making?* The proposed study will address this question in two parts (*see Figure 4*). First, I evaluate correlations of teacher assigned grades and standardized assessments and explore whether or not the correlations have strengthened over time. I argue that a strong positive correlation could increase the potential that district leaders and teachers

could use grades as valid data measures of achievement in decision making (*Figure 4, left column*). Second, I examine student grade patterns to cluster students and understand how past student grades predict future student outcomes. Specifically, I hope to pinpoint specific times and subjects for early instructional intervention for specific students (*Figure 4, right column*).

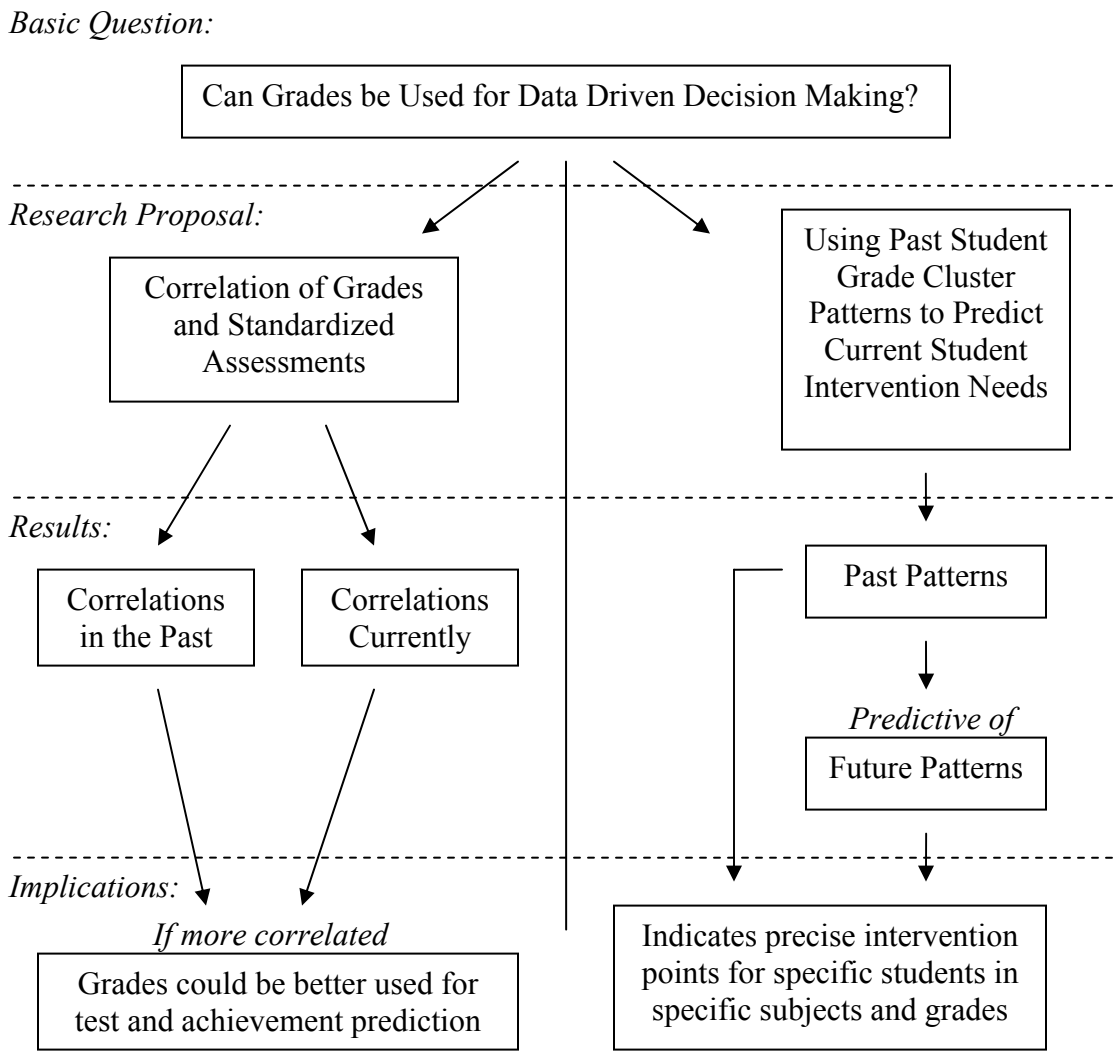


Figure 4: *General Flow of the Dissertation Framework.*
The proposal of this dissertation is to study if grades can be better used for data driven decision making in K-12 schools. The general flow of the proposed questions and data is presented, with hypothesized results and general implications.

Research Questions

1. To what extent has the correlation between grades and standardized assessments changed from earlier student cohorts to more recent cohorts?
2. To what extent does the Hargris hypothesis of past grading patterns predicting future student grade patterns hold true?
3. To what extent is grade patterning useful in predicting student outcomes such as graduation or dropping out? To what extent do these predictive patterns aid in identifying avenues for early intervention by instructional leaders and teachers?

CHAPTER IV: METHOD

Sample

For this study, the entire assessment histories of a sample of students were collected, including grades and standardized assessments. The sample of students was comprised of all of the students of the entire graduating cohorts of 2006 and 1994 (whether or not they graduated) for two districts, West Oak and South Pine (pseudonyms). Districts were selected based on their comparative small sizes (less than 3000 students each) to keep the study at a reasonable size for a single researcher to complete data collection over a three month time period, their relative diversity in student populations, and their willingness to participate in the study. Both districts are located in the American Mid-West, are located within 20 miles of each other, and are first ring suburbs of a large metropolitan area. In addition, both districts are currently undergoing dramatic demographic changes as their populations shift from a majority European American demographic, to an increasing population of Hispanic and African American families. For issues of confidentiality, district specifics are intentionally left vague.

West Oak is defined as a mid-sized central city by the U.S. census, with less than 3000 students attending two elementary schools, a middle school and a high school. In 2006, the district served a student population that was about 70% economically disadvantaged, 50% Hispanic, 30% European American and 15% African American. The district has historically lagged behind the state averages on state standardized tests in both reading and mathematics at all grade levels (NCES, 2006; S&P, 2006).

South Pine is defined as an urban fringe of a mid-sized city by the U.S. census, with fewer than 3000 students attending three elementary schools, a middle school and a

high school. In 2006, the district served a student population that was about 50% economically disadvantaged, 50% European American, 20% Hispanic, and 15% African American. The district has historically scored near the state averages on state standardized tests in both reading and mathematics at all grade levels (NCES, 2006; S&P, 2006).

Data Collection

Students were included in the sample if they started first grade with the student cohort expected to graduate from high school in either May of 1994, or 2006. For both districts, the first grade school year was 1982/1983 for the graduating class of 1994, and school year 1994/1995 for the graduating class of 2006. Two cohorts were selected in each of the two districts to provide an initial comparison of grading and standardized assessments over time. The 2006 cohort was selected as the most recently graduated cohort from each district. The 1994 cohort was selected because it was the oldest cohort in West Oak for which student data files contained both grading histories and state standardized test score records. For comparison, the graduating cohort of 1994 was also included for South Pine. Thus, four cohorts of students comprise the sample.

Each student's permanent record in paper form was accessed from the district's long-term paper file storage. Student data was entered into SPSS, using a unique identifier to de-identify each student. No student names were recorded for this study. For each student, grades for every subject for every year were recorded, K through 12. Additionally, scores for each standardized test on record were recorded, including composite and subject specific scores. Standardized tests included the state standardized tests for grades 3, 4, 5, 6, 7, 8, and 10 as well as the ACT. Because it was outside of the

scope of this study, attendance was not recorded. For students who transferred into the district on track to graduate in either 1994 or 2006, if the student's file contained grades and assessment data from their past school district, those grades were also recorded.

For high school grades, both the letter grade and the name of each class taken were recorded for each semester for grade 9 through 12 to provide a rich dataset in which both the subject grades were recorded as well as the name of each subject-level class for each semester for each grade level. Classes were grouped by the following subjects; Mathematics, English, Science, Foreign Language, Social Science, Economics, Band, Physical Education/Health, Computers, Life Skills, Family Skills, and Art. Accordingly, multiple class grades of the same subject but for different classes were recorded within the same subject and grade level variable, so that for each student one column of data was recorded as the different class names for a subject during a specific grade level and semester, and the next column was the letter grade for that subject in that grade level and semester. As an example, the data recorded for first semester 10^{th} grade mathematics class name for all cohorts included classes such as Algebra, Geometry, Trigonometry, and Math Skills, among others. The letter grade for each student for each of these different classes was recorded under the variable name "Math Grade 10 Semester 1". This was repeated for all high school classes.

Because the two districts over the two time periods recorded Middle School and elementary grades differently by semester, some recording just the final yearly grade and some recording by semester, and also because some of the schools had different semester schedules in which the 180 day school calendar was divided into 2, 3 or 4 semesters, all Middle School and elementary grades were recorded as "composite grades". To generate

the composite grade, letter grades for each subject for each semester recorded were first converted to the following numeric grading scale: A=4.0, A- = 3.666, B+ = 3.333, B = 3.0, B- = 2.666, C+ = 2.333, C = 2.0, C- = 1.666, D+ = 1.333, D = 1.0, D- = 0.666, E or F = 0. Then, the mean grade for that school year was calculated from the numeric grades to generate the composite grade. Composite grades were then entered into SPSS similarly to the high school grades by subject.

Although course names at the elementary level were fairly consistent across districts, time periods and report cards, early elementary grading marks were not. This posed an interesting dilemma as to how to record subject specific grades for each student at each grade level for grades K (kindergarten) through 3. Table 1 presents the different grade marking scales identified from the various report cards for grades K through 3. Interestingly, while few report cards for these grades used the more standard A,B,C,D grading scale, all conformed to some form of a four point scale. No matter the scale used, from pluses and checks, to V, S, N, O, to 1,2,3,4, teachers awarded students based on a four point scale that mirrored the classic A,B,C,D scale. Interestingly, except for the one report card in the sample that used the symbol grading scale, teachers commonly used the +/- designations to represent a degree of achievement between scoring ranks. As examples, with the VSNO scale, V^- or S^+ was a common designation, or with the 1,2,3 scale a 1^- or a 2^+, indicating a mark between the top mark and next highest mark (*Table 1*). With the grading scales mirroring the traditional four-point scale, each grading period's mark by subject was converted to a numeric grade according to the scheme presented in Table 1. The mean for all of the grading periods for each subject in a specific grade level was then recorded.

Table 1: *A Four-Point Grading Scale and the Differential Grading Marks of Elementary Teachers, Grades K-3*

Standard	A	A^-	B^+	B	B^-	C^+	C	C^-	D
Check	+	$+^-$	$\surd^+$	$\surd$	$\surd^-$	$-^+$	–	$-^-$	O
VSN	V	V^-	S^+	S	S^-	N^+	N	N^-	O
OSN	O	O^-	S^+	S	S^-	N^+	N	N^-	SE
123	1	1^-	2^+	2	2^-	3^+	3	3^-	4
ABCN	A	A^-	B^+	B	B^-	C^+	C	C^-	N
ABPH	A	A^-	B^+	B	B^-	P^+	P	P^-	H
Symbol	^			O			X		┌
Numeric Conversion	4	3.6	3.3	3	2.6	2.3	2	1.6	1

Symbol Key:
V – Very good, S – Satisfactory, N – Needs Improvement
1 – Excellent progress, 2 – Progressing at expected level, 3 – Needs to improve, 4 – Special needs
^ - Demonstrates effectively, O – Demonstrates some, X – Working, ┌ - Does not demonstrate
P – Progressing, H – Help needed, SE – See comments
Note: "O" was used differently for multiple scales

Additional variables were also recorded for each student, including gender, date of birth, ethnicity, and student transfer status, both in and out of the district. The issue of the designation of "dropout" is highly contested in the literature (Greene & Winters, 2005; NCES, 2004; Swanson, 2004; Viadero, 2006) and official definitions differ by state and by region. Nevertheless, many students who were on track to graduate on-time with their cohort in this sample did not. Because the term "dropout" is currently under contention in the literature and policy domains, for this study, as has been previously recommended (Ensminger & Slusarcick, 1992; Marrow, 1986), students were designated as either On Time Graduation – students who had evidence of receiving a diploma on-time with their cohort or had evidence of a valid transfer out of the district – or Not On Time Graduation (NOTG).

A student was considered to have graduated on-time if their record contained evidence of the award of a diploma. A valid student transfer was defined as any student's

record which contained a request for student transcripts from another school district or school which was not an alternative school. Although a student who transferred to another district may have eventually dropped out, there is no way to determine this, and as has been previously recommended (Ensminger & Slusarcick, 1992; Marrow, 1986), valid transfer students are designated as on-time graduators.

A record of a transcript request from an alternative school was defined as a non-valid indicator of student transfer for on-time high school graduation, and thus was an indicator of the educational challenges faced by the student with a high probability that the student would not graduate on-time with their cohort. Lacking confirming graduation or alternative degree completion data from the alternative education schools, it can not be determined if the students who transferred to alternative education programs graduated on-time with their cohort with a full high school diploma, rather than a G.E.D. It is the case that many students who transferred to alternative high schools had low or failing grades in multiple subjects at the time of the transfer. Past research on the G.E.D. option has shown that it is not equivalent to a regular high school diploma (Cameron & Heckman, 1993; Tyler, 2003) and thus is not considered for this study as on-time graduation with a standard high school diploma. Even if these students did graduate from an alternative high school with a diploma or an alternative high school degree (G.E.D.), this study is focused on the on-time graduation of the cohort of students in a traditional high school program, and so thus will consider students who transferred to an alternative education program as NOTG. Interestingly, it has been shown previously that students identified as "at risk" for dropping out are often directed to an alternative education program by district personnel before they drop out (Sipple *et al.*, 2004), and that the

inclusion of a GED option may encourage students to drop out (Tyler, 2003). If a student's file did not contain a record of a diploma award, a request for student records from another district, or the record ended prematurely, that student was designated as not on time graduation (NOTG). Thus, NOTG should be considered as a "proxy" for student dropout that may contain some unknown degree of false positives; students who are categorized as NOTG but did graduate on time.

Statistical Analysis

All data entry and statistical analyses, except for cluster analysis (see below) and calculation of confidence intervals between correlations (see Appendix B), were conducted using the statistical software package SPSS 14 (SPSS, 2006). For the purposes of statistical analysis in this study, subject specific grades are considered as ordinal variables while GPA, state standardized test scale scores and ACT subject specific and composite scores are considered as interval scales. Thus, when correlating subject specific grades to other measures, a nonparametric statistic, Spearman's Rho, is utilized. However, when correlating interval measures, Pearson product moment correlations are used where indicated (Howell, 2002). To calculate confidence intervals between two independent correlations, Fischer's r to z transformation and ρ was utilized (Howell, 2002) (see Appendix B).

Cluster Analysis

To test the grade patterning hypothesis of the ability of past student grade patterns to predict future student outcomes based on current student grades, cluster analysis was used to identify the underlying patterns within the K-12 grading dataset. To supply the

clustering procedure with ample data as well as the subsequent analysis of the results, the entire grading histories recorded within the dataset were included where appropriate.

Cluster analysis is a descriptive statistical analysis that brings empirically defined organization to a set of previously unorganized data (Anderberg, 1973; Eisen *et al.*, 1998; Jain & Dubes, 1988; Lorr, 1983; Rencher, 2002; Romesburg, 1984; Sneath & Sokal, 1973). There are two types of clustering, supervised and unsupervised. Supervised clustering begins with a defined set of assumptions about the categorization of the data, while unsupervised clustering assumes nothing about the categorization and is designed to statistically discover the underlying structure patterns within the dataset (Kohonen, 1997), a procedure well suited to discovering the underlying patterns within student grades. While there are many types of unstructured cluster analyses (Anderberg, 1973; Lorr, 1983; Romesburg, 1984; Sneath & Sokal, 1973), this study will focus on hierarchical cluster analysis, due to the procedure's ability to discover a taxonomic structure within a dataset efficiently (Lorr, 1983; Rencher, 2002; Romesburg, 1984; Wightman, 1993).

Hierarchical clustering provides a way of organizing cases based on how similar the values for the list of variables are for each case. In hierarchical clustering, each case is first defined as an individual cluster, a series of numbers for each variable on that case. As an example, this could be a single student's grades in all subjects K-12. Then, at each level of clustering, the two most similar cases are joined based on how similar the pattern of numbers is for both cases, as defined by a similarity distance measure, discussed below. This continues in a hierarchical fashion as similar cases are joined to clusters and clusters are themselves joined to similar clusters, until the clustering algorithm defines

the entire dataset at the highest hierarchical level as one cluster (Anderberg, 1973; Eisen et al., 1998; Lorr, 1983; Rencher, 2002; Romesburg, 1984; Sneath & Sokal, 1973). Thus, when complete, cases that were previously organized just as a pseudo-random descriptive list, organized alphabetically or by student numbers, are placed nearby other cases in the list with which they have a high similarity, aiding in visualization and identification of empirically defined patterns previously unknown within the dataset. To date, while few studies in education use clustering, those that have describe their clustering results in many varied ways (Sireci *et al.*, 1999; Wightman, 1993; S. Young & Shaw, 1999). Descriptions range from unintuitive, to verbose, to difficult to interpret. One way to help visualize the organization of the data by hierarchical clustering is to draw a cluster tree, sometimes referred to as a dendrogram (Eisen et al., 1998; Lorr, 1983; Romesburg, 1984). Within a cluster tree, clusters of cases and clusters of clusters can quickly be identified by the closeness of lines corresponding to cases and linked to other cases. The unit length of the line indicates similarity of patterns, the distance in the data space between the two clusters in the units of the measure, with a shorter line denoting higher similarity.

Recently, researchers in the biological sciences, specifically molecular biology, where the human genome project has produced massive amounts of data, have made innovations in using and visualizing hierarchical clustering. Confronted with unordered and unintuitive displays of datasets that include tens of thousands of genes with thousands of data points for each gene in multiple samples, traditional techniques are unworkable. One quickly adopted innovation was the Eisenplot. First invented by Michael Eisen at Stanford, the Eisenplot takes tables of clustered numbers, which the

human mind can not easily interpret for pattern recognition, and converts the table into blocks of color, aiding the human eye in visualizing patterns within clustered data (Eisen et al., 1998). In addition, while traditional statistical program packages do include clustering algorithms, such as SAS (using PROC CLUSTER) and SPSS, due to the explosion of genetic data and the near ubiquitous use of hierarchical clustering by molecular biologists, clustering programs and visualization software are now also freely available on the internet (DeHoon *et al.*, 2004; Eisen, 1998; Eisen & DeHoon, 2002; Vilo, 2003).

Because cluster analysis has been rarely used in educational research, a simplified example of the clustering procedure is informative to detail the process, and the algorithms used. To allow for initial visualization in two dimensional space, and to keep the example relatively brief, the example data set includes five fictitious students, numbered 1 through 5, with 8th grade English and Mathematics grades (*Table 2*).

Table 2: *Example 8th Grade Dataset, English and Mathematics for Letter Grades for 5 students*

Student ID	English Grade	Mathematics Grade
1	D	D
2	B+	A
3	A	C
4	A	C+
5	D	C

Visual inspection of the letter grading patterns between the five students is difficult, as the list is ordered only by the student number. Imagine if the data set were to include the data of an entire cohort with grades in many more subjects for multiple years. Discerning patterns in the grading data would be impossible without the aid of a clustering method.

The example clustering method detailed here is an adapted version of Romesburg's overview of cluster analysis (1984). I substitute subject specific grading data into this analysis, along with the addition of an Eisenplot as the final step in cluster visualization, which is not discussed by Romesburg.

Hierarchical cluster analysis for use with grade data first requires that student letter grades be converted to a four point scale (*for the conversion scheme, see Table 1*). For the example hypothetical data, the letter grades data in Table 2 are converted to numeric grades data in Table 3.

Table 3: *Example 8th Grade Dataset, English and Mathematics Numeric Grades for 5 students*

Student ID	English Grade	Mathematics Grade
1	1.0	1.0
2	3.6	4.0
3	4.0	2.0
4	4.0	2.3
5	1.0	2.0

This dataset can be visualized in two dimensions using a scatter plot (*Figure 5*).

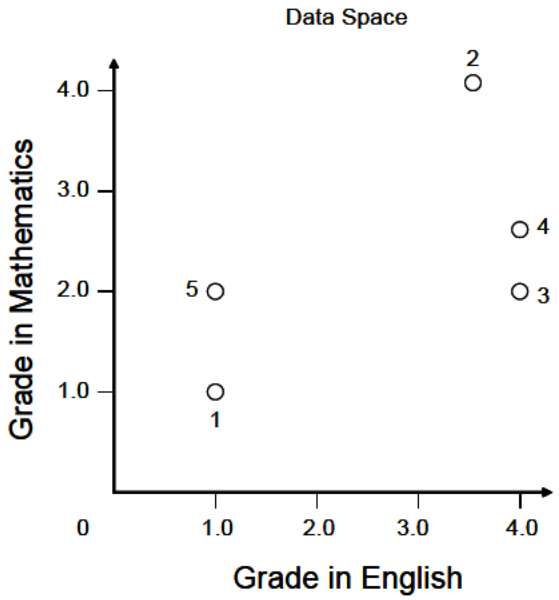


Figure 5: *Scatter plot of example data set*

In the Figure 5 scatter plot, each student is identified by student ID, and each student's numeric grade (*Table 3*) is plotted in the two dimensional "data space", which is an intersection of the English and Mathematics grade data. From this plot, the proximity of each student's data pattern in the data space can be visualized. If the dataset were to include hundreds of student cases, rather than five, and hundreds of subject specific grades, rather than two, visualization in this manner would quickly become impossible as the number of cases would overlap and the number of dimensions of data would surpass three, and thus become impossible to visualize. Cluster analysis, and the resulting visualization techniques, enables the quantification of case proximity within a multi-dimensional data space as a measure of similarity between cases, as well as dendrogram and Eisenplots to aid in pattern recognition and visualization of the clustering results.

In cluster analysis, the goal is to determine quantitatively how similar each case's data pattern is to every other case, cluster the two most similar cases' data patterns, and

repeat the process until the entire dataset is defined as a single cluster, thus determining the hierarchical data structure within the data. This is accomplished through the following eight steps:

1. Create a resemblance matrix by calculating a distance measure between every case.
2. Combine the two most similar cases into a cluster.
3. Use a clustering algorithm to recalculate the resemblance matrix.
4. Iterate over steps 2 and 3 until all of the cases are clustered into one cluster, e.g. n-1 times.
5. Rearrange the order of the cases on the basis of their similarity according to the results of step 4.
6. Draw the dendrogram.
7. Draw the Eisenplot.
8. Interpret the clusters.

For step 1, a distance measure between every point in Figure 5 must be calculated. This can be accomplished through a variety of methods (Anderberg, 1973; Lorr, 1983; Rencher, 2002; Romesburg, 1984; Sneath & Sokal, 1973). To present a simplified example, the Euclidean distance will be used for the hypothetical dataset. The distance between each case can be represented as a dashed line, as shown in Figure 6.

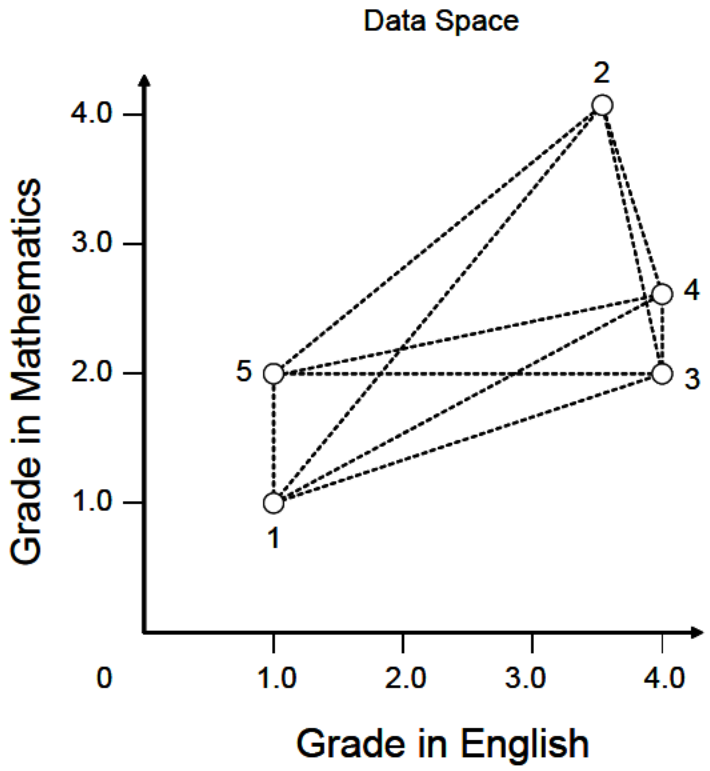


Figure 6: *Right triangles drawn between each example data point in the data space*

The length of each line is a measure of the similarity of each case to each other case in the two dimensional grade data space. To calculate the length of each line, the generalized Pythagorean theorem (Euclidean distance) can be used in which each line is considered the hypotenuse of a right triangle and the length of the hypotenuse is determined through the formula $a^2+b^2=c^2$. The more general form of this equation, the Euclidean distance, for any two series of numbers in which $x = \{ x_1, x_2, \ldots, x_n \}$ and $y = \{ y_1, y_2, \ldots, y_n \}$ is defined as:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Equation 1

Using Equation 1, a lower left triangular resemblance matrix can be generated for the example data, as shown in Table 4.

Table 4: *Example data resemblance matrix, step 1, iteration 1*

	Student ID				
Student ID	1	2	3	4	5
1	*				
2	4.01	*			
3	3.16	2.03	*		
4	3.28	1.70	0.33	*	
5	1.00	3.33	3.00	3.02	*

Each cell in Table 4 is the calculated Euclidean distance, using Equation 1, between each of the five students in the data space in grading units. At this point, each student is considered a cluster, and in the subsequent steps, will be grouped into larger clusters with students who have similar data patterns.

The second step is to combine the two cases with the shortest distance into a new cluster. The smallest distance measure in the example is between student 3 and student 4, 0.33 (*Table 4*). This is intuitive from the relative distance seen between these two cases in Figures 5, 6 and 7. Students 3 and 4 are "closest" in the data space, and so should be the first two cases to cluster together. In this fashion, the first cluster is defined as cluster 34, and Figure 6 can be updated to show this cluster, as in Figure 7.

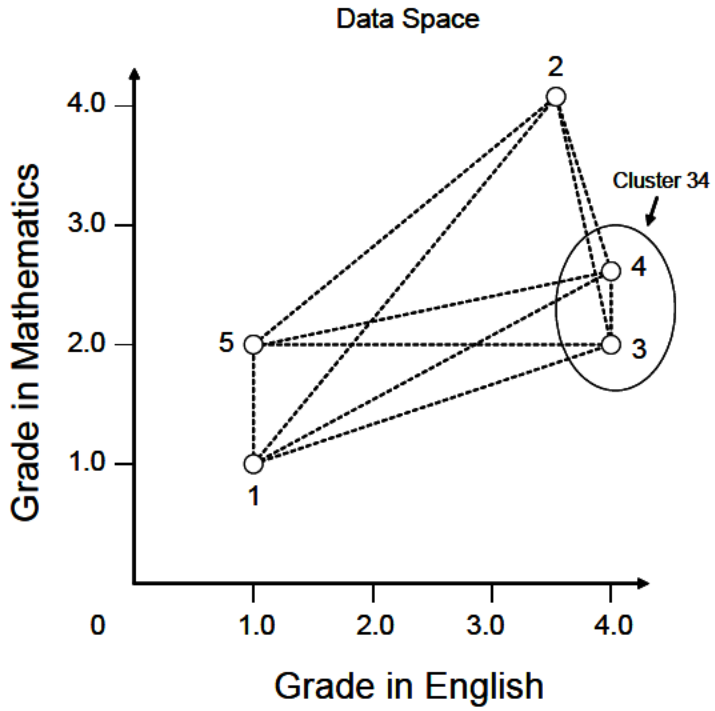


Figure 7: *Example data set, cluster 34 defined*

The third step is to employ the use of a clustering algorithm. Many algorithms have been suggested in the literature, however the *average linkage* method is known to provide good results and is accepted as a standard clustering algorithm (Eisen & DeHoon, 2002; Lorr, 1983; Rencher, 2002; Romesburg, 1984; Sneath & Sokal, 1973). It is the clustering algorithm utilized for this study. For average linkage, if d(x,y) is equal to Equation 1, the distance between any two clusters A and B is defined as the average distance of the total number of cases within both clusters, $n_A n_B$, between the total number of cases in cluster A, n_A, and the total number of cases in cluster B, n_B, such that:

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{i=1}^{n_B} d(x_i, y_i)$$

Equation 2

where the sum is over all of x_i in A and all of y_i in B. For each step of the clustering, the two clusters with the smallest distance are joined and the resemblance matrix is recomputed according to Equation 2. For the example then, the updated resemblance matrix is shown in Table 5.

Table 5: *Example data resemblance matrix, step 3, iteration 2*

	Cluster ID			
Cluster ID	1	2	5	34
1	*			
2	4.01	*		
5	1.00	3.33	*	
34	3.222547	1.863914	3.009212	*

The two clusters with the smallest Euclidean distance as calculated using Equation 2 are cases 1 and 5. These two cases are then combined into cluster 15. The data space may now be represented as in Figure 8.

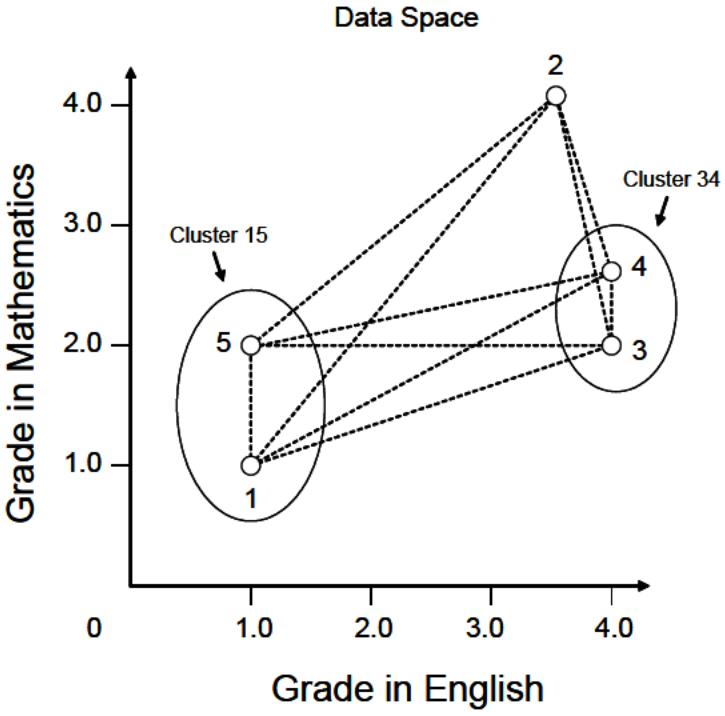


Figure 8: *Example data set, cluster 34 and 15 defined*

Repeating step 4, the resemblance matrix is recalculated using Equation 2. The matrix results are shown in Table 6.

Table 6: *Example data resemblance matrix, step 3, iteration 3*

	Cluster ID		
Cluster ID	2	34	15
2	*		
34	1.86	*	
15	3.67	3.12	*

From Table 6, the two most similar clusters are clusters 2, and 34 with a distance of 1.86. These two clusters are combined into a new cluster, 234. The data space can now be represented as in Figure 9.

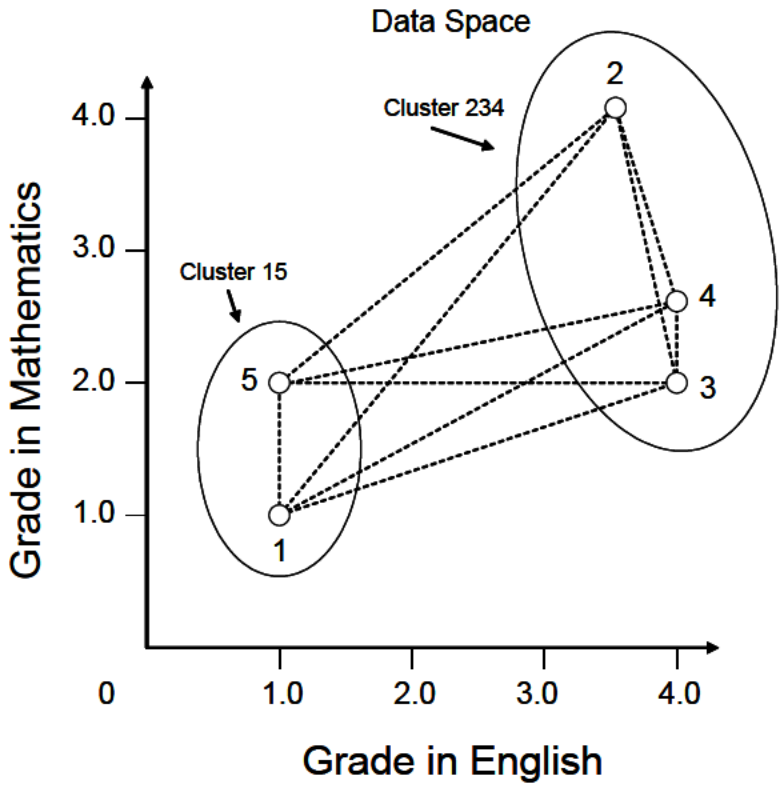


Figure 9: *Example dataset, cluster 234 and 15 defined*

The final iteration of the clustering algorithm to update the resemblance matrix gives the data in Table 7, and the entire example data set is included in the final cluster, 12345.

Table 7: *Example data resemblance matrix, step 3, iteration 4*

	Cluster ID	
Cluster ID	15	234
15 *		
234	3.30 *	

Therefore, through these steps, the unordered list of student IDs can now be reordered by the similarity of each example student's grade pattern in 8th grade English and mathematics, as 1, 5, 2, 3, and 4. Cluster analysis also provides a means to visualize this order, and the relative magnitude of the difference or similarity between clusters, known as a dendrogram, or a cluster tree (Eisen et al., 1998; Rencher, 2002; Romesburg, 1984), as shown in Figure 10 for the example data set.

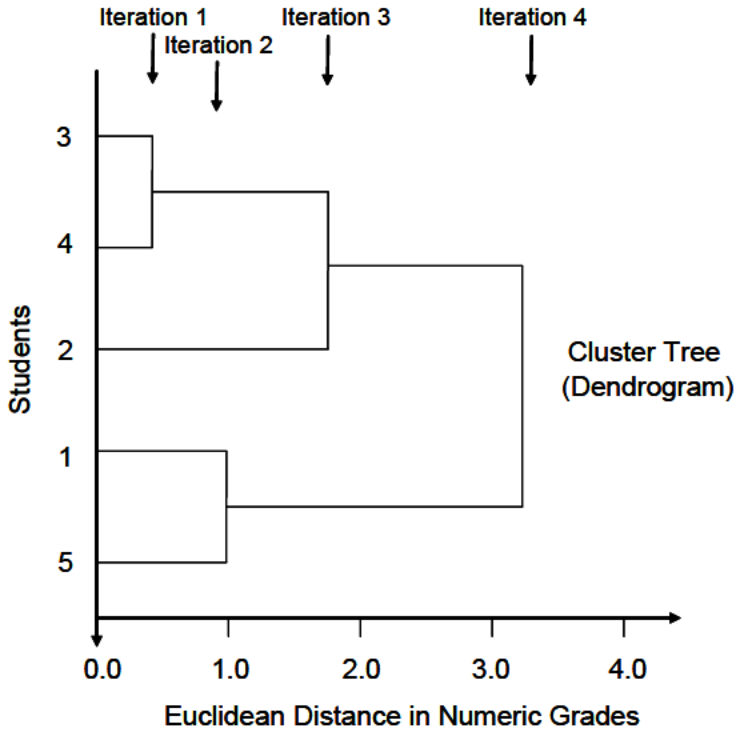


Figure 10: *Example dataset dendrogram*

The cluster tree "grows" with each iteration of the clustering algorithm. The x-axis represents the Euclidean distance between each data point in the data space in numeric grade units. The y-axis represents each student cluster. The length of each horizontal line in the tree is defined by the distance measure calculated at that iteration for that cluster. So, for the most similar cluster, 34, which was defined during the first iteration, the distance is 0.33 (*Table 4*). The next most similar cluster is cluster 15, with a distance of 1.0 (*Table 5*). Clusters 34 and 2 are linked in the tree at height 1.86 because these two clusters were the next most similar at iteration 3 (*Table 6*). The final height of the tree is defined by the final calculation in the resemblance matrix in iteration 4, 3.30 (*Table 7*). Thus, the dendrogram allows for the visualization of the order and magnitude of the similarity of each student, based on the clustering of each student's grade pattern within the multi-dimensional data space.

Step 7 is a more recent addition to cluster visualization, the inclusion of an Eisenplot, pioneered in molecular biology and cancer research (Bowers *et al.*, 2000; Eisen et al., 1998; van'tVeer et al., 2002; Weinstein *et al.*, 1997). In this step, each student's grade patterns are converted into blocks of color, aiding the human eye in pattern identification across multiple cases and multiple data patterns (Eisen et al., 1998; Weinstein *et al.*, 1997). Thus, multiple **images in this dissertation are presented in color**. In addition, categorical data that may be informative in interpreting clustering patterns can be visualized along with each case's data pattern. For the example data, adding a hypothetical categorical variable such as "on time graduation" to the data presented in Table 3 would result in the data show in Table 8.

Table 8: *Example 8th grade dataset, English and mathematics numeric grades and one categorical variable*

Student ID	English Grade	Mathematics Grade	On Time Grad
1	1.0	1.0	0
2	3.6	4.0	1
3	4.0	2.0	1
4	4.0	2.3	1
5	1.0	2.0	0

The clustered data can then be represented, along with the categorical variables, in a manner that allows for visualization of the clusters, as well as the data patterns of each case and the relation of categorical variables to the cluster patterns.

As suggested by Eisen and others, an Eisenplot should display cases as rows and data categories as columns, such as subject specific grades (Eisen et al., 1998; van'tVeer et al., 2002; Weinstein et al., 1997). Each data point is represented by varying intensities of color blocks, according to a heat-map. For this study, the heat-map will range from a deep red for the highest scores, to a grey for the middle scores, to a deep blue for the lowest scores, (*Figure 11, scale*). An Eisenplot for the example fictitious data presented in Table 8 is shown in Figure 11.

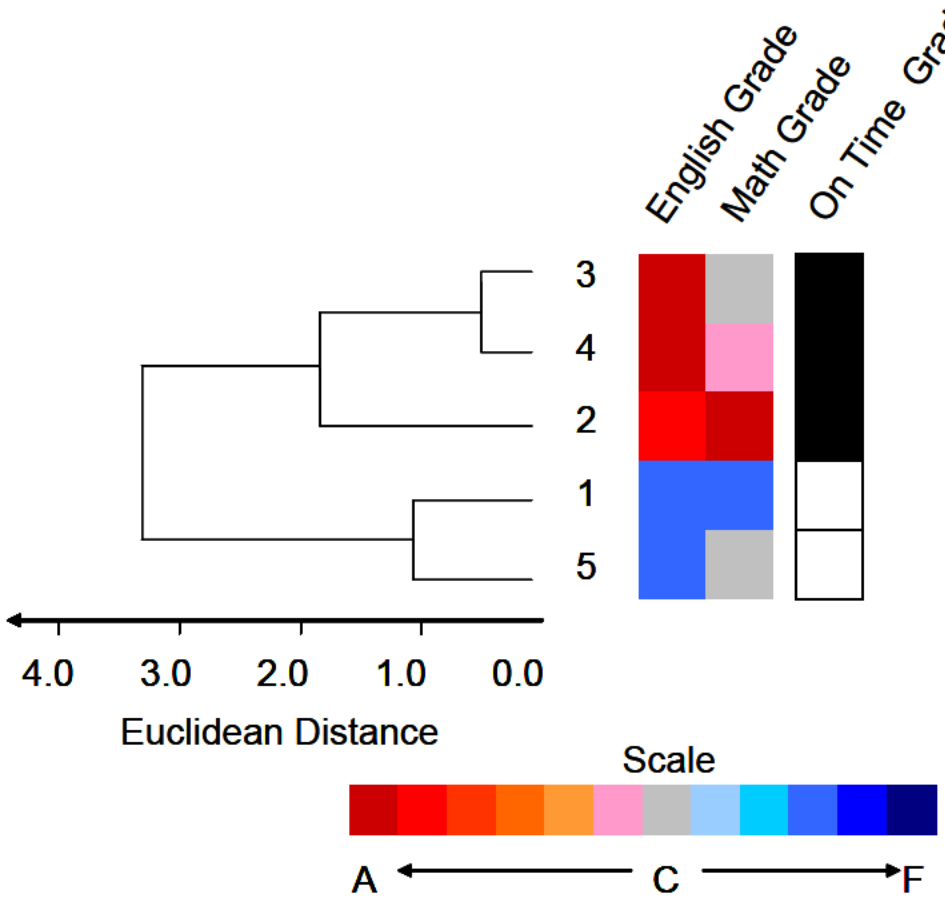


Figure 11: *Dendrogram and Eisenplot of the example dataset*

Figure 11 combines all of the data from Table 8 as well as the cluster analysis presented in Figure 9, into a single figure which shows the hierarchical clustering structure of the data with a dendrogram and a cluster-ordered list of cases (*Figure 11, left*), a color-coded representation of the data patterns for each case (*Figure 11, center color blocks*), and a representation of the categorical variables in which a black block indicates the presence of the on time graduation variable, and a white block the absence (*Figure 11, right*). Hence, an Eisenplot is a figure which allows for cluster analysis interpretation through the presentation of all of the data in the entire data set ordered through hierarchical clustering. Figure 11 shows that for the example data set, students 3

and 4 were the most similar in their English and Mathematics grades followed by student 1 and 5. Student 2 was the most dissimilar of the data set, but was more similar to cluster 34, than to cluster 15 (*Figure 11, left*). Cluster 234 scored higher than cluster 15 overall in English and Mathematics (*Figure 11, color blocks*), and for this fictitious example data set, on time graduation was associated with generally higher grades in English and Mathematics (*Figure 11, right*). Thus, this example has shown how the use of hierarchical cluster analysis can order an unordered list of cases based on the similarity of the data patterns of those cases, and display that information in an interpretable and intuitive data display. However, the example presented above is a simplification of the cluster analysis method used in this study, namely through the use of Euclidean distance, and was presented in this primer because the Pythagorean Theorem is readily understood and produces easily interpretable results.

The hierarchical clustering strategy employed in this study differs from the above example in two ways, standardization of scores and the use of uncentered correlation as the distance measure. Overall, the steps of the clustering method parallel the above detailed method. First, the data matrix Y was obtained which contained the data for all four cohorts of students with every subject specific grade, K-12:

$$Y = \begin{pmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_n \end{pmatrix}$$

Equation 3

in which y'_i is an observation vector corresponding to each student case, and $y_{(j)}$ is a column corresponding to subject specific numeric grades, converted from letter grades as detailed above. Second, each $y_{(j)}$ was normalized through z-scoring, so that the data in the entire matrix Y was replaced with z-scores based on the means of each subject specific and grade-level specific column, $y_{(j)}$. This step is recommended to control for overweighting in the clustering algorithm by arbitrary cases (Rencher, 2002; Romesburg, 1984). Third, publicly available online clustering software was used to cluster the data (Vilo, 2003). The distance measure employed was uncentered correlation, which differs from the above hypothetical example. A correlation based measure has been recommended in the literature as superior to Euclidean distance (Rencher, 2002) and is commonly used in hierarchical clustering (Eisen & DeHoon, 2002). The most commonly used correlation based measure is the Pearson product moment correlation coefficient, in which for any two series of numbers $x = \{ x_1, x_2, \ldots, x_n \}$ and $y = \{ y_1, y_2, \ldots, y_n \}$

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

Equation 4

The Pearson product moment correlation is where $\bar{x}$ is the mean of the values of series x, $\bar{y}$ is the mean of the values of series y, σ_x is the standard deviation of series x, and σ_y is the standard deviation of series y. However, a modified version of the Pearson product moment correlation is known as uncentered correlation and is defined as:

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_x^{(0)}}\right)\left(\frac{y_i}{\sigma_y^{(0)}}\right)$$

Equation 5

in which

$$\sigma_x^{(0)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i)^2}$$

Equation 6

$$\sigma_y^{(0)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i)^2}$$

Equation 7

The function defined in Equation 5 is highly similar to the Pearson correlation in Equation 4, except that it assumes that the mean is 0 for every series even when it is not. This is important when considering two vectors, x and y, that have the same shape but are separated by a constant value. The Pearson correlation (a centered correlation) would be the same for these two vectors, namely 1, while the uncentered correlation for these two vectors would not be 1 (Anderberg, 1973; Eisen & DeHoon, 2002). Stated in different terms, "the uncentered correlation is equal to the cosine of the angle of two n-dimensional vectors x and y, each representing a vector in n-dimensional space that passes through the origin" (Eisen & DeHoon, 2002, p.11). It is this uncentered correlation which was used to calculate the distance measure for the hierarchical clustering in this study. This required the use of a modified parametric statistic, uncentered correlation,

with semi-ordinal data, grades. This is an appropriate distance measure for this data based on the recommendations of the current data mining and bioinformatics literature. Furthermore, it should be noted that the choice of which distance measure is "best" for any particular application is currently under contention (Anderberg, 1973; Ein-Dor *et al.*, 2006; Eisen & DeHoon, 2002; Eisen et al., 1998; Jain & Dubes, 1988; Lorr, 1983; Lu *et al.*, 2005; Romesburg, 1984; Shen *et al.*, 2006; Sneath & Sokal, 1973; vandeVijver *et al.*, 2002; Weinstein et al., 1997; Zapala & Schork, 2006). Hence, while the question of which clustering algorithms perform best with subject specific grades is of interest, it is outside the scope of this study. Additionally, as in the example presented above, the average linkage clustering algorithm was also employed in this study. Cluster dendrogram and Eisenplots were generated as detailed in the example above using publicly available online clustering software (Vilo, 2003).

CHAPTER V: GRADES AND STANDARDIZED ASSESSMENTS

Description of Sample

The sample for this study consisted of two entire student cohorts, for two school districts, West Oak and South Pine, and included all students who entered either district on-track for graduation with their cohort in either 1994 or 2006. The overall descriptive variables for the sample are presented in Table 9.

Table 9: *Descriptive variables and frequencies for all students in the sample*

Overall Study Descriptive Variables	
Total Number of Students Sampled	361
Percent NOTG[§]	26.3
Percent with IEPs	15.2
Gender (%)	
Female	49.6
Male	50.4
Ethnicity (%)	
European American	58.7
Hispanic	12.5
African American	6.1
Asian	1.9
Multi-ethnic	2.0

§ Excludes West Oak 1994 cohort due to lack of non-graduating student data

From Table 9, overall, 361 students were included in the sample. Females and males were almost evenly split, while the ethnic majority of the sample is European American, followed by Hispanic, African American, multi-ethnic and Asian students. Out of all four cohorts included in the sample, 15.2% of the students had at least one year in which an individual education plan (IEP) was included in the student's file, indicating that the student had been recommended for special education services at some point throughout their time within the district. The overall graduation rate for the sample was 72.9%, and

thus the NOTG (Not On Time Graduation) was 26.3%. However, for the 1994 cohort in West Oak, unfortunately the district had purged its files at some point in the past of all non-graduating students, and thus did not have any student data files for the 1994 cohort of students who did not graduate. Hence, the overall graduation rate and NOTG percentages are most likely not valid indicators of the overall sample on-time graduation rates when the West Oak 1994 cohort is included.

To understand better the student demographics of each cohort for each district, student demographic variables are disaggregated in Table 10 by district and year.

Table 10: *Descriptive variables and frequencies by district and cohort year*

	West Oak		South Pine	
Descriptive Variables	1994	2006	1994	2006
Total Number of Students Sampled	36	105	130	90
NOTG	---[§]	34.3	36.9	12.2
Percent with IEPs	27.8	17.1	13.1	10.0
Gender (%)				
Female	44.4	41.0	50.8	60.0
Male	55.6	59.0	49.2	40.0
Ethnicity (%)				
European American	83.3	28.6	73.8	62.2
Hispanic	8.3	29.5	2.3	8.9
African American	0	9.5	2.3	10.0
Asian	2.8	0	1.5	4.4
Multi-ethnic	0	1	0.8	5.5
No Ethnicity Data	5.6	31.4	0	8.9

[§] Excludes West Oak 1994 cohort due to lack of non-graduating student data

Due to the vagaries of district data collection and retention, while many student's records included data such as ethnicity, for both districts, multiple students did not have any ethnicity recorded. This issue with missing ethnicity data was most prevalent for the West Oak 2006 cohort, with 31.4% of the student records containing no information on

ethnicity (*Table 10*). In addition, the data for the West Oak 1994 cohort includes only those students who graduated on time, as described above.

From Table 10, it is obvious that both school districts are under dramatic demographic ethnic shifts, from a European American majority to a more diverse student cohort, including many more Hispanic and African American students. Additionally, both districts have experienced a decrease in the number of students with records of IEPs from 1994 to 2006, from 27.8% to 17.1 for West Oak, and 13.1% to 10% for South Pine. However, for West Oak, this data is difficult to interpret due to the lack of NOTG student data.

The demographic shift of both communities is made more obvious when the United States Census Bureau estimates of demographic populations for both the 1990 and 2000 census are considered ("U.S. Census bureau", 2007). For West Oak, while the overall population was stable, in 1990 94% of the population was ethnically European American, 4% Hispanic, 1% African American and 1% Asian. In 2000 for West Oak, the percentages changed to 73% European American, 20% Hispanic, 5% African American and 2% Asian. For South Pine, a similar trend occurred. The population of South Pine grew by 13% between 1990 and 2000. In 1990, 94% of the population was European American, 2% Hispanic, 2% African American and 1% Asian. In 2000, for South Pine, the percentages shifted to 86% European American, 6% Hispanic, 5% African American, and 3% Asian. While the 1990 and 2000 community census data does not directly parallel the 1994 and 2006 cohorts by time and sample, it is obvious that the student populations and the communities of both districts are experiencing demographic shifts over time.

Standardized Assessments and Grades

To begin to address the hypothesis of standardized assessments and teacher assigned grades converging over time, standardized assessments and subject-specific grades were collected for each student in the sample. Standardized assessments included: the ACT (ACT, 2007), generally taken by a subset of each cohort sometime during the 11th grade academic year; the state's standardized assessment in multiple subjects given in grades 3, 6, 8 and 10; and subject specific grades for all grade levels K-12. A brief summary of the assessment data, overall and by cohort, is given in Table 11. ACT composite scores are measured on a scale from 1 to 36, Grade Point Average (GPA) is measured on a four-point scale as discussed in the methods. State test scores were measured on a four-point category scale according to the following scheme: 4 – "not endorsed"; 3 – "endorsed at basic level"; 2 – "endorsed met state standards"; 1 – "endorsed exceeded state standards".

Table 11: *Means for assessment data for the full dataset, and by cohort*

		West Oak		South Pine	
Assessment	Overall	1994§†	2006	1994†	2006
ACT Composite Score	19.984	22.00	18.61	20.36	20.05
	(4.342)	(4.950)	(4.149)	(4.114)	(4.167)
% of Cohort who took the ACT	34.6	47.2	34.3	25.4	43.3
10th Grade State Test‡					
Mathematics	2.636	--	2.811	--	2.513
	(0.901)		(0.941)		(0.856)
Science	2.636	--	2.818	--	2.500
	(0.838)		(0.862)		(0.798)
Social Studies	3.053	--	3.200	--	2.947
	(0.844)		(0.890)		(0.798)
Reading	2.331	--	2.444	--	2.246
	(0.618)		(0.664)		(0.572)
Writing	2.522	--	2.589	--	2.474
	(0.622)		(0.626)		(0.618)
High School GPA	2.347	2.641	2.311	2.070	2.626
	(0.909)	(0.670)	(0.849)	(0.984)	(0.832)
High School GPA by Subject					
Math	2.035	2.361	1.848	1.947	2.184
	(1.016)	(0.897)	(0.997)	(1.050)	(0.995)
English	2.252	2.570	2.323	1.960	2.455
	(1.048)	(0.939)	(1.036)	(1.041)	(1.029)
Science	2.098	2.049	2.027	1.960	2.359
	(1.032)	(0.853)	(1.065)	(1.091)	(0.954)

Note: Standard deviations are presented in parentheses below each mean
§ West Oak 1994 sample only includes students who graduated on-time
† 1994 state assessment scores not comparable to 2006
‡ State test scores reported by proficiency categories

Examining the data in Table 11 in more detail, ACT composite scores decreased significantly for West Oak between the 1994 and 2006 cohorts, $t(51)=2.61$, $p<0.05$, but not for South Pine, $t(70)=0.32$, $p=0.751$. Overall high school GPA decreased significantly for West Oak between the 1994 and 2006 cohorts, $t(114)=2.06$, $p<0.05$. Conversely,

overall high school GPA increased significantly for South Pine between the 1994 and 2006 cohorts, $t(207)=-4.32$, $p<0.001$. Similar trends were also observed for subject specific GPAs. Overall, these data suggest that the 1994 West Oak cohort performed better than South Pine on the ACT, but because West Oak has seen a decrease in composite ACT scores between the 1994 and 2006 cohorts, while South Pine has remained stable, South Pine's 2006 cohort appears to be outperforming the West Oak 1994 cohort on the ACT. If this difference is truly a trend in the data attributable to the actions of each district, rather than to exogenous variables such as cohort effects, it is especially interesting given the similar demographic shift that each district is currently experiencing. With ethnically changing populations, West Oak has seen declines in ACT scores, while South Pine has maintained stable ACT scores. It would be interesting to continue to track these trends between the two districts.

It should again be noted that the 1994 West Oak cohort did not include any NOTG students, so comparisons of grades between 1994 and 2006 for West Oak is problematic. However, as will be described below in chapter VI, almost all of the students who took the ACT graduated on-time for the three cohorts which contain NOTG student data. It will be assumed for this study that the same was true for the West Oak 1994 cohort, so that while overall GPA may not be comparable for West Oak from 1994 to 2006, ACT scores are comparable since it appears that students who take the ACT generally graduate on-time, and so are included in the 1994 West Oak sample.

Unfortunately during data collection, it was found that the state standardized test data would not be comparable between the 1994 and 2006 cohorts. The state in which both districts are located has undergone multiple rounds of state assessment design over

the past two decades, especially for the high school assessments, such that both the test itself and the reporting methods for the test have changed dramatically. Test scores recorded in each student's file for the 1994 cohort were not reported as either category scores nor scale scores, and thus were not comparable to the test scores reported for the 2006 cohort, which were reported as category scores and scale scores. Thus, one initial finding of this study is that for any districts within the state studied, state test scores are not comparable over the twelve year time span for the 1994 and 2006 student cohorts, because scores are not on equivalent or matched scales.

While the initial hope to use state standardized test scores to compare to grades over time could not be realized, the districts for all four cohorts did record a comparable standardized assessment for multiple subjects, namely the ACT. The percentages of each cohort which took the ACT are presented in Table 11. South Pine has seen a dramatic increase in the percentage of students taking the ACT from the 1994 cohort to 2006. For West Oak, the difference in percentages can not be interpreted since the 1994 West Oak cohort does not contain any NOTG students. Thus, since state standardized test scores can not be used to compare the 1994 and 2006 cohorts, ACT scores will be used instead to examine the hypothesis of if grades and standardized assessments are converging over time. Using ACT scores is not ideal, since less than half of each cohort took the ACT, and almost none of the students who took the ACT were NOTG. The assessment scores of this large majority of students who did not take the ACT can not be determined using the data collected. While this study will now turn to comparing ACT scores and grades, the results will be applicable to only the students who took the ACT in each cohort and also were enrolled in the district and had grades recorded.

To further explore the ACT data presented in Table 11, the subject specific ACT scores for each cohort are presented in Figure 12. For all subsequent figures which refer to the subject specific ACT tests, the following abbreviations will be used: MATH – mathematics subtest; ENG – English subtest; READ – reading subtest; SCI – science subtest. Boxplots were constructed for each ACT subject specific subtest for each cohort and were compared (*Figure 12*).

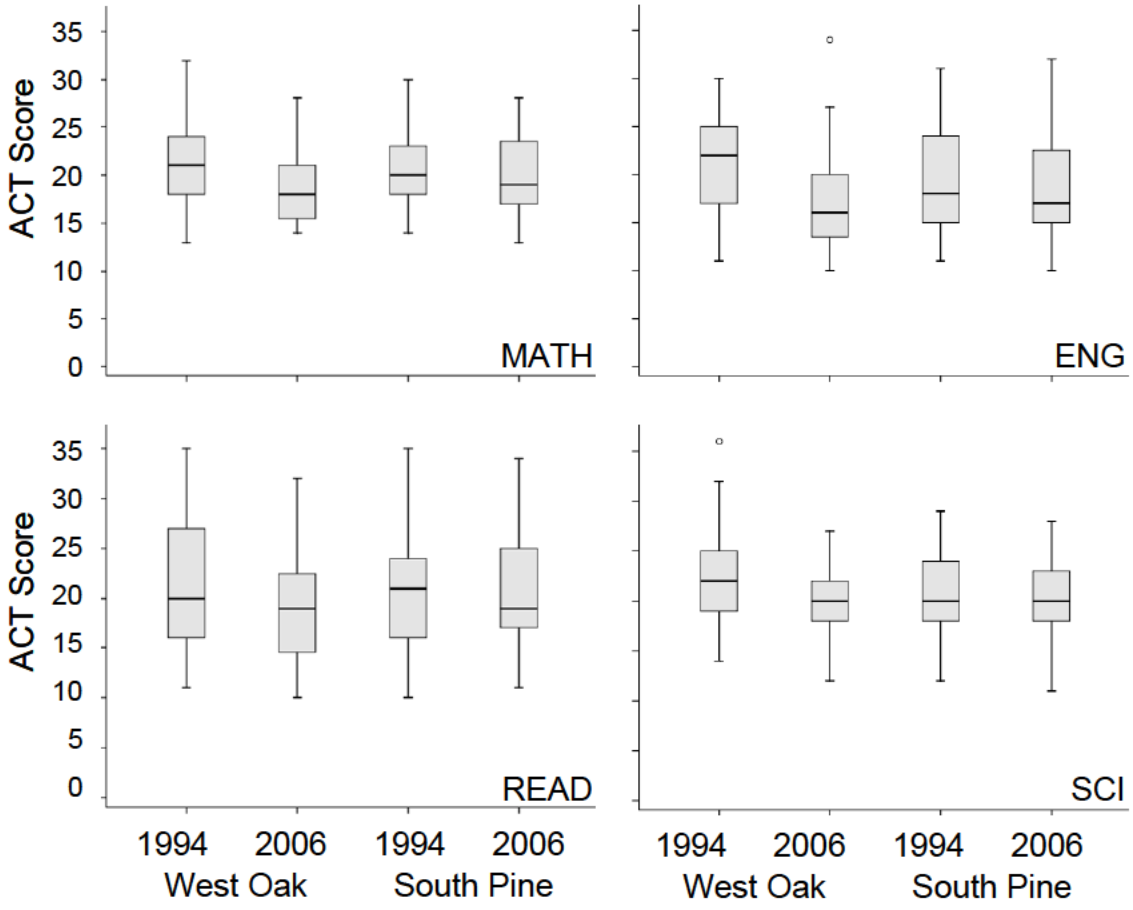


Figure 12: *Boxplots of ACT subject-specific subtest scores by cohort*

Figure 12 is a set of standard boxplots in which the center line represents the median value for each group, the lower and upper boundary for each box represents the border of the first and third quintiles respectively, the whiskers above and below represent the 1.5 value interquartile range beyond the box, and the circles represent outliers that are beyond the 1.5 interquartile range. For the ACT subtest data, the differences between the districts and cohorts are similar to the overall ACT composite averages (*compare Table 11 and Figure 12*). Specifically, while the median ACT score for all four subtests has decreased in West Oak between the 1994 and 2006 cohorts, and remained relatively stable in South Pine, the ACT scores for the 2006 West Oak cohort are somewhat less variable than all of the other scores in each subject (*second set of box and whiskers from the right, all four panels, Figure 12*). Overall variability in ACT subtest scores appears to be less for all four cohorts in mathematics and science, while it is the highest in reading. The subtest data appear to be generally normally distributed with few outliers, other than West Oak 2006 mathematics, both South Pine cohorts in English, and the South Pine 2006 cohort in reading (*Figure 12*). Thus, the ACT subtest data is generally symmetric and appears generally normally distributed, and parallels the overall trends of the ACT composite means.

Historically, student scores on a subject specific subtest of a standardized test, such as the ACT, are highly correlated with the other subject specific subtests on that same test (Brennan *et al.*, 2001; Linn, 1982; Woodruff & Ziomek, 2004) due to test design, student ability, student knowledge and test-wiseness (Mehrens & Lehmann, 1991). In this study, ACT composite and subject scores are highly correlated for the full dataset, replicating the previous research (*Table 12*).

Table 12: *Correlations of ACT composite and subject subtest scores, full dataset*

Subject Test	*Composite*	*ENG*	*MATH*	*READ*	*SCI*
Composite	1.0				
ENG	0.883***	1.0			
MATH	0.830***	0.642***	1.0		
READ	0.910***	0.776***	0.647***	1.0	
SCI	0.842***	0.599***	0.725***	0.709***	1.0

Note: Correlations are Pearson product moment correlations, and $n = 124$ for all correlations
*** $p<0.001$

As seen in Table 12 for the full dataset, ACT subject subtests in English, mathematics, reading and science all highly correlate with the overall composite score, each exceeding a correlation of 0.8. Correlations between each subtest were also high, but to a lesser extent than with the composite score. Specifically, the lowest subtest correlation was between English and science, followed by English and mathematics. The highest subtest correlation was between English and reading (*Table 12*). These results are not surprising given the subject matter of the subtests, in that it is intuitive that English and reading would highly correlate whereas English and mathematics might not correlate as highly. The lower correlation between English and science is interesting, since reading skill is a component of science instruction (Yore *et al.*, 2003).

The comparison of a standardized assessment, such as the ACT, to teacher assigned grades is of interest for multiple research contexts (Brennan et al., 2001; Girotto & Peterson, 1999; Linn, 1982; Woodruff & Ziomek, 2004) including the current study with a focus on the possible convergence of grades and standardized assessment systems over time. However, historically in comparing ACT scores and grades, actual grades are

rarely used. Rather, the ACT corporation collects survey information from participating students and asks students to self-report their grades in multiple subjects (Woodruff & Ziomek, 2004). While interesting, student self reported grades are a problematic source of information on actual student grades and recently have been critiqued as not accurately reflecting actual grades in the subjects surveyed (Kuncel *et al.*, 2005). The current study helps to address this important issue and add to the research literature by using actual recorded teacher assigned grades in comparison to ACT scores for every student who took the ACT for two cohorts in two districts each, one of the first studies to do so.

Teacher assigned subject specific grades were recorded as detailed in the methods. To simplify this discussion, the following comparisons to ACT scores will utilize overall high school GPA, subject-specific GPAs, as well as focusing on 10th grade second semester subject-specific grades. Because the ACT was taken sometime during the 11th grade academic year for the majority of students in the sample, to explore how subject-specific grades correlate with ACT scores, 10th grade semester 2 grades provide a set of teacher assigned subject-specific grades that were awarded to students prior to the year in which they took the ACT. These grades should reflect the cumulative ability of each student in each subject as judged by their teachers, taking into account all of the issues of hodge-podge and subjective grading detailed in chapter II. However, to help address the issue of individual teacher bias during 10th grade semester 2, overall GPA and subject-specific GPA are also compared to ACT scores below.

Subject-specific grades were recorded for each class taken by each student, and classes were categorized into subjects for each high school semester and grade level. Grades were then grouped by subject for each semester and grade level. Subject grouping

categories were as follows: mathematics, English, science, foreign language, social studies, government, economics, band/music, physical education/health, computers, life skills, art. The distribution of the types of classes taken during 10th grade semester 2 across the full dataset is shown in Figure 13.

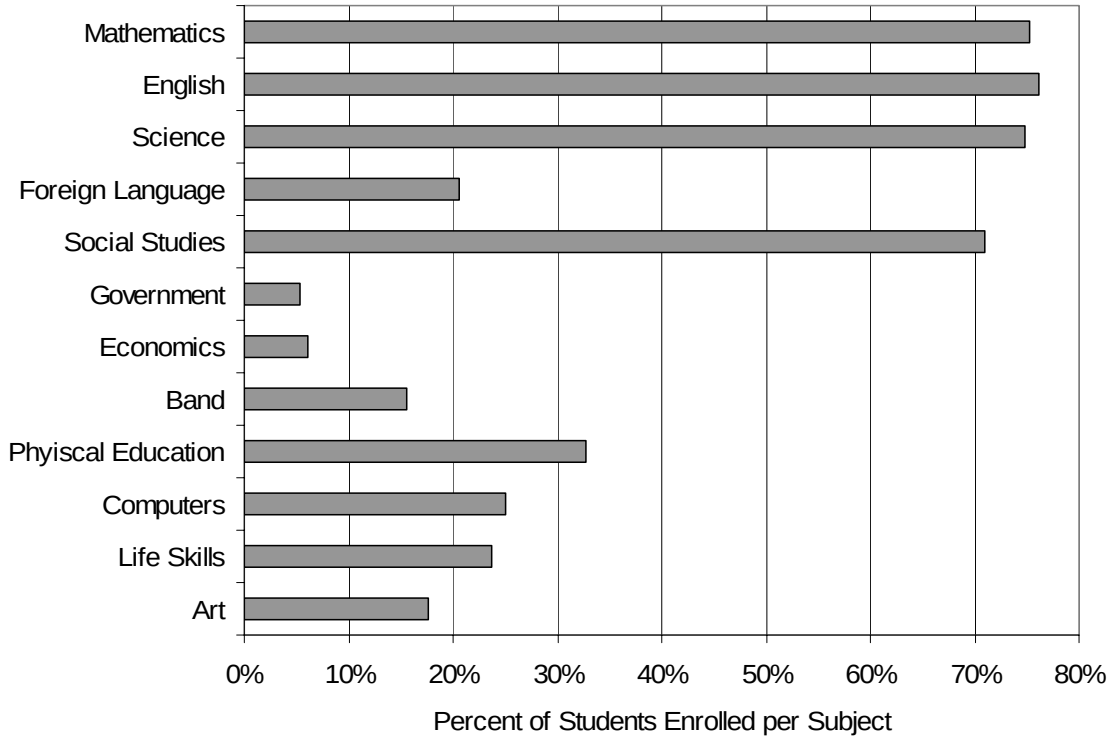


Figure 13: *Distribution of the types of classes taken during 10th grade semester 2, full dataset*

For the full dataset, during 10th grade semester 2, over 70% of students took a core set of classes that focused on mathematics, English, science and social studies (*Figure 13*). For the remaining subjects, over 20% of the students took a class that dealt with a foreign language, physical education, computer or life skills. Classes that focused on government, economics, band or art all enrolled less than 20% of the students in the dataset. The course names and percentages of students who were enrolled in each specific course for each subject grouping during 10th grade semester 2 are detailed in Appendix A.

As with the correlations above of each subject-specific subtest of the ACT with the other subtests of the ACT, it is of interest to examine how well grades in each subject correlate with each of the other subjects for 10th grade semester 2. Grades for each subject in 10th grade semester 2 for the entire dataset were correlated with the grades in each of the other subjects using Spearman's Rho correlations (*Table 13*). For the core set of classes taken by most of the students including mathematics, English, science and social studies, highly significant correlations are above 0.5, many above 0.6. The highest correlation among the core set of classes is between science and social studies at 0.709, the lowest is between mathematics and social studies at 0.536. Among the other subjects with an *n* over 30, many of the correlations also appear fairly high, ranging from about 0.4 to 0.7. Subjects such as foreign language and computers appear to correlate at about 0.5 across the core set of classes of mathematics, English, science and social studies. Interestingly, band, physical education, life skills, and art do not appear to correlate as highly with the core set of classes, ranging in correlations from about 0.3 to 0.5. These differences in correlation may be interpreted in several ways. First, because mathematics, English, science and social studies are considered a core curriculum and are tested by the state test as well as the ACT, teacher grading practices across these subjects may be more aligned than with non-core subjects such as band, physical education, life skills and art. Additionally, the curriculum and grading practices of subjects such as foreign language and computers may be more aligned with the core set of classes than with the non-core.

Table 13: *Correlations of subject-specific grades for 10th grade semester 2, full dataset (Spearman's Rho)*

Subject	*Math*	*English*	*Science*	*Foreign Language*	*Social Studies*	*Government*	*Economics*	*Band*	*Physical Education*	*Computers*	*Life Skills*	*Art*
Math	1.0											
English	0.599*** (263)	1.0										
Science	0.636*** (257)	0.619*** (262)	1.0									
Foreign Language	0.644*** (72)	0.735*** (72)	0.626*** (69)	1.0								
Social Studies	0.536*** (243)	0.587*** (247)	0.709*** (244)	0.643*** (65)	1.0							
Government	0.477* (19)	0.715** (19)	0.661** (18)	0.738 (4)	0.233 (9)	1.0						
Economics	0.740*** (21)	0.763*** (20)	0.598** (22)	1.000 (2)	0.682** (18)	0.500 (3)	1.0					
Band	0.353* (47)	0.410** (50)	0.428** (52)	0.725** (14)	0.419** (50)	0.354 (5)	0.500 (3)	1.0				
Physical Education	0.314** (105)	0.466*** (107)	0.312** (102)	0.154 (22)	0.411*** (96)	0.529 (9)	-0.500 (3)	0.162 (19)	1.0			
Computers	0.528*** (83)	0.440*** (84)	0.527*** (87)	0.662** (25)	0.596*** (81)	0.500 (3)	0.277 (6)	0.338 (15)	0.576** (30)	1.0		
Life Skills	0.429*** (88)	0.437*** (88)	0.398*** (86)	0.457 (14)	0.331** (83)	-- (1)	0.125 (5)	0.364 (9)	0.542** (22)	0.285 (27)	1.0	
Art	0.535** (63)	0.459*** (64)	0.330** (61)	0.407 (19)	0.355** (58)	0.866 (3)	1.000 (2)	0.521 (8)	0.589** (24)	0.445 (16)	0.521* (17)	1.0

Note: The *n* of each correlation is in parentheses below the correlation
*** p-value < 0 001
** p-value < 0 01
* p-value < 0 05

A second interpretation may be that student performance in the core subjects is similar to their performance in each of the other core subjects. One can imagine that the lessons learned on how to negotiate the grading system may be similar in subjects that require similar types of participation, homework and assessments, such as mathematics, English, science and social studies. However, for subjects that may require a different set of skills to demonstrate achievement, participation, homework and assessment, such as band, physical education, life skills, and art, student performance may not correlate as well with the core set of subjects or with any of the other non-core set of subjects (*Table 13*). Unfortunately, due to low *n* with few of the same students taking classes across the non-core subjects, correlations with subjects such as government and economics, as well as correlations between the non-core subjects are difficult to interpret. It would be of interest for future studies to delve further into these differences, collecting a larger sample, to show if across a broader population of students the correlation of grades remains fairly high for the core subjects, and lower for the non-core subjects.

Despite these differences in correlations, Table 13 shows that, for the full dataset, student grade performance is similar across core subjects, with significant correlations over 0.5, and also is somewhat similar with non-core subjects, with significant correlations over 0.3. Historically, since grading data has been thought of as difficult to collect, few studies have dealt with the correlation of subject specific grades, often lacking the data altogether and relying on GPA or self-reported grades (Kuncel et al., 2005). Of the studies that have collected subject specific grades, almost all sample a population of students, rather than collect entire cohorts of data (Alexander *et al.*, 2001; Brennan et al., 2001; Girotto & Peterson, 1999). However, as with the correlations of

ACT subject subtests, and as shown here in Table 13 for the full dataset, grades in one core subject correlate with grades in other core subjects (Brennan et al., 2001). This correlation may be the result of any of the interpretations discussed above, from teacher and curriculum assessment alignment, to student acquired skill at negotiating the hodge-podge grading system and knowing how to participate, hand-in homework and show up for class (Brookhart, 1991; Cross & Frary, 1999). In addition, the high correlation of core subject grades may also be due to student aptitude (Jencks & Phillips, 1999) in which student innate ability in core subjects influences the grade teachers assign. However, no matter the interpretation of the correlations, for the data presented for the full dataset, if a student's grade is high or low in one core subject, that same student's grade in another core subject is likely to be very similar, as is the case with correlations of ACT subtest scores. This implies that for the core subjects of mathematics, English, science and social studies, student grades and ACT test performance depend more on the student than on the specific subject, in that student achievement appears to be somewhat subject independent. This result implies that to examine the main hypothesis for this chapter of the correlation of grades and standardized assessments, it would be advantageous to examine achievement scores in multiple subjects for both grades and ACT simultaneously, rather than just GPA or ACT composite scores, to further explore this student cross-subject performance result as well as to show if the correlation between grades and ACT scores over time has changed.

Additionally, when 10th grade semester 2 subject-specific grades are correlated with the ACT composite and subtest scores for the full dataset, core subject scores show moderate and significant correlations (*Table 14*), replicating past research which has

shown similar moderate correlations between grades and standardized assessments (Brennan et al., 2001; Linn, 1982; Woodruff & Ziomek, 2004). Interestingly, as opposed to the moderate intra-subject correlations for 10th grade semester 2 grades between core subjects and non-core subjects, such as mathematics and art (*Table 13*), correlations between 10th grade semester 2 grades for non-core subjects and the ACT composite and subtests are lower, and mostly not statistically significant (*Table 14*). This may be due to the low *n* for the number of students in the full dataset who took non-core classes and the ACT.

Table 14: *Correlations of ACT composite and subtest scores with 10th grade semester 2 grades, full dataset (Spearman's Rho)*

Subject Grades, 10th Grade Semester 2	*ACT Composite*	*ACT MATH*	*ACT ENG*	*ACT READ*	*ACT SCI*
Mathematics	0.398*** (121)	0.517*** (120)	0.345*** (120)	0.289** (120)	0.294** (120)
English	0.578*** (123)	0.423*** (122)	0.510*** (122)	0.557*** (122)	0.458*** (122)
Science	0.441*** (121)	0.451*** (120)	0.334*** (120)	0.420*** (120)	0.370*** (120)
Foreign Language	0.458** (47)	0.262 (47)	0.466** (47)	0.366* (47)	0.373** (47)
Social Science	0.548*** (113)	0.499*** (112)	0.378*** (112)	0.515*** (112)	0.502*** (112)
Government	0.705* (10)	0.375 (10)	0.773* (10)	0.452 (10)	-0.003 (10)
Economics	-0.051 (5)	-0.872 (5)	0.158 (5)	0.359 (5)	-0.296 (5)
Band	0.347 (28)	0.171 (28)	0.282 (28)	0.485** (28)	0.263 (28)

Physical Education	0.151 (45)	0.063 (44)	-0.061 (44)	0.159 (44)	0.253 (44)
Computers	0.296 (42)	0.391* (42)	0.234 (42)	0.289 (42)	0.314* (42)
Life Skills	0.275 (28)	0.117 (28)	-0.143 (28)	0.355 (28)	0.457* (28)
Art	0.096 (29)	0.199 (28)	-0.016 (28)	0.069 (28)	0.013 (28)

Note: Correlations are Spearman's Rho
*** $p<0.001$
** $p<0.01$
* $p<0.05$

However, this difference in correlation between the correlation of core and non-core subject grades versus the correlation with ACT subtests (*compare Tables 13 and Table 14*) may also be due to the difference between what is measured by grades versus what is measured by the ACT, in that while the ACT may measure the acquisition of knowledge, grades also measure the acquisition of knowledge (because they correlated with the ACT) but also may measure a student's success at negotiating the social processes of schooling and the hodge-podge subjective grading system. This would hypothetically result in a moderate correlation between core and non-core subject grades, which is seen in this study (*Table 13*). For this study, this type of correspondence between core and non-core subject grades is termed a "success at school factor" (SSF), in which the similar variance between the grades in two or more different subjects may be attributable to a student's ability at negotiating school as an overall social process, while the non-similarity is attributed to the differences in the correlation between core and non-core subjects. The moderate correlation between ACT and core subjects (*Table 14*) is attributable to the similar knowledge needed for both assessments, but because the ACT

does not correlate with non-core subjects (*Table 14*) while core subject grades do moderately correlate with non-core subjects (*Table 13*). As one possibility, this may suggest that the ACT measures only one part of grades – knowledge acquisition – which is about 25% of the variance in grades (0.5 correlation of ACT and core subject grades), while grades may measure knowledge acquisition plus another variable that is not related to what is measured by the ACT. This result, in combination with the above finding that student grade performance appears to be somewhat subject independent, leads to the hypothesis here that this other variable is a "Success at School Factor" (SSF). The evidence presented here is admittedly initial evidence only, with the major threat to the validity of this argument coming directly from the small and intact samples that are biased towards students who take the ACT. This topic of a possible Success at School Factor will be further addressed in chapters VI and VII.

One way in which to test a success at school factor would be to correlate subject-specific grades with a standardized test given to a broader population than the ACT was given. This would help to include students not included in the above tables, such as students who do not graduate on time or who chose not to pursue college. This point can be tested with this dataset using standardized state high school test scores for the 2006 cohorts for both West Oak and South Pine. While the use of the state standardized test narrows the student sample to just the two 2006 cohorts, it broadens the type of student included, since the vast majority of students took the state standardized high school tests, both on-time graduators and NOTG students. This analysis rests on the assumption that the standardized state test is similar to the ACT, in that it assesses the extent of student academic knowledge. If the state test correlates with core subject grades only, there is

support for the hypothesis that teacher assigned grades may assess both academic knowledge and a success at school factor. These correlations are presented in Table 15.

Table 15: *Correlations of standardized state high school test scale scores with 10th grade semester 2 grades, 2006 cohorts – West Oak and South Pine (Spearman's Rho)*

	State Standardized Tests				
Subject Grades, 10th Grade Semester 2	*Math*	*Science*	*Social Studies*	*Reading*	*Writing*
Mathematics	0.399***	0.357***	0.291**	0.272**	0.161
	(122)	(120)	(122)	(118)	(124)
English	0.373***	0.485***	0.397***	0.452***	0.325***
	(121)	(119)	(121)	(117)	(123)
Science	0.392***	0.467***	0.496***	0.444***	0.256**
	(121)	(119)	(121)	(117)	(123)
Foreign Language	0.282	0.482**	0.304	0.354*	0.517***
	(38)	(36)	(39)	(36)	(40)
Social Studies	0.482***	0.566***	0.543***	0.466***	0.262**
	(112)	(111)	(114)	(108)	(114)
Government	0.498	0.758**	0.467	0.543	0.523
	(13)	(12)	(12)	(13)	(13)
Economics	-0.026	0.410	0.821	0.821	0.553
	(5)	(5)	(5)	(5)	(5)
Band	0.253	0.084	0.156	0.303	0.102
	(26)	(26)	(25)	(26)	(26)
Physical Education	0.227	0.196	0.189	0.214	0.338**
	(57)	(56)	(58)	(55)	(59)
Computers	0.253	0.212	0.223	0.369**	0.356**
	(54)	(54)	(55)	(53)	(56)
Life Skills	0.496**	0.349	0.234	0.332	-0.157
	(28)	(29)	(28)	(28)	(28)
Art	0.068	0.119	0.083	0.214	0.185
	(47)	(46)	(46)	(46)	(47)

Note: Correlations are Spearman's Rho
*** $p<0.001$
** $p<0.01$
* $p<0.05$

The data presented in Table 15 supplies further evidence supporting the possible existence of a SSF component in grades. Grades from 10th grade semester 2 core subjects, such as mathematics, English, science and social studies, moderately correlate with the

state standardized test across multiple subject tests, including mathematics, science, social studies, reading and writing (*Table 15*). However, the state subject tests generally do not correlate with non-core subject grades, such as band, physical education and art. Also, Table 15 supplies additional evidence that both grades and the standardized state test scores are independent of the actual subject and appear tied more to each individual student. The moderate correlations suggest that students who do well or do poorly in one subject will generally have similar scores across the other subjects assessed. These data supply an initial test of the success at school factor, and indicate that grades may measure two important factors in the lives of students, academic knowledge and success at school. If this is true, these results have important implications for the emphasis on standardized tests as the main driver of data driven decision making in schools, districts, states and the nation. As will be detailed in chapter VI, student's grades can predict if a student will or will not graduate on-time. Acknowledging that graduation from high school is an important predictor of student life outcomes, investing in a better understanding of a possible success at school factor and its possible assessment through grades, and non-assessment through standardized tests, has deep implications for school leaders engaged in data driven decision making. These issues will be further taken up in chapter VII.

The correlation of grades and standardized assessments

The hypothesis for this chapter is that grades and standardized assessments may be converging over time, as discussed in chapters II and III. To explore this hypothesis, correlations of grades and ACT scores between the 1994 and 2006 cohorts for both districts are examined in this section. First, correlations of ACT scores with overall high school GPA for both years for both districts will be presented, then the more detailed

subject-specific high school GPAs, followed by the fine grained 10th grade semester 2 subject specific grades.

The Pearson product moment correlations of ACT composite and subject-specific subtest scores with overall high school GPA (HSGPA) varies dramatically across the two cohorts in each of the two districts (*Figure 14*). For West Oak, the correlations of 1994 (*Figure 14, left panel, dashed line*) and 2006 HSGPA (*Figure 14, left panel solid line*) with ACT scores are highly similar for both years for the ACT composite, reading and science subtests, but vary dramatically and inversely for the mathematics and English ACT subtests. The high variation for the 1994 West Oak cohort may be due to the small sample size. The South Pine correlations are somewhat more moderated, in that the correlations of 1994 (*Figure 14, right panel dashed line*) and 2006 HSGPA (*Figure 14, right panel solid line*) with ACT scores are relatively similar for the ACT composite, mathematics and science subtests, but differ for the English and reading subtests. Overall, the correlation of HSGPA the ACT shows a mixed results with both districts showing little change between the 1994 and 2006 cohorts for the correlation of HGSPA and ACT composite, and multiple differences across the ACT subtests. Confidence intervals for these correlations across each cohort for each district all overlap (*see Appendix B*), indicating that there is no statistical difference in the correlation between HSGPA and ACT scores for both cohorts in both districts.

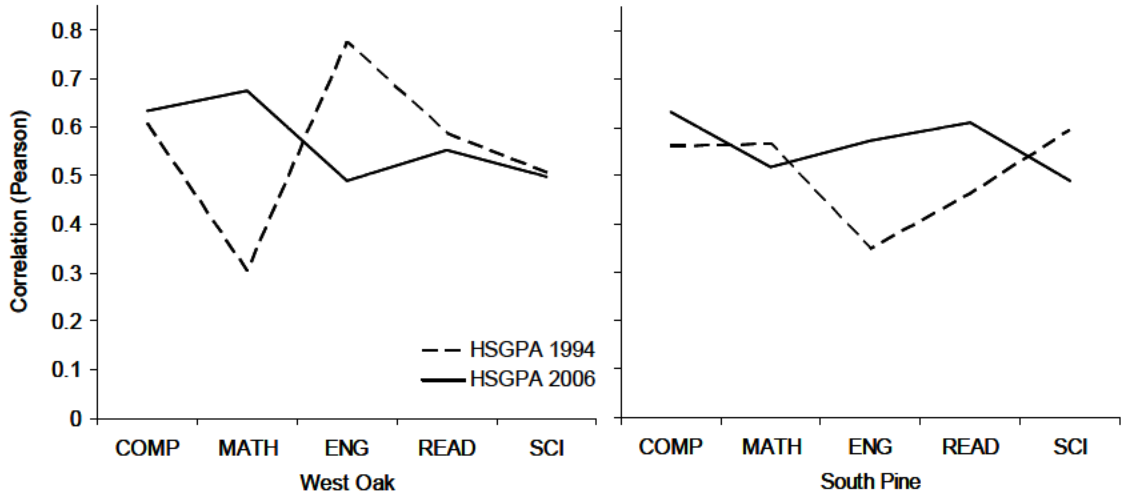


Figure 14: *Correlation of high school GPA and ACT between the 1994 and 2006 cohorts for both West Oak and South Pine (Pearson correlations)*

In addressing the question of if the correlation of grades and ACT scores are converging over time, the data in Figure 14 answers the question to some extent. For both districts, the correlation between HSGPA and ACT has increased slightly from the 1994 cohort to the 2006 cohort, but the difference is exceedingly small and not statistically significant. Additionally, the differences between ACT subtest scores and HSGPA correlations may indicate that the ACT composite or the HSGPA is masking trends that are occurring in less aggregated data. In addition, from the data presented above in Table 14, ACT scores do not correlate well with non-core subject grades that an overall measure of grades, such as high school GPA, includes. If grades and standardized assessments are converging over time, that convergence may only be occurring in core subjects, especially core subjects that are tested by standardized tests such as the ACT, namely mathematics, English and science. Thus, to delve deeper into the change in correlations over time, subject-specific GPAs in mathematics, English and science were correlated with the ACT subtest scores (*Figure 15 and Figure 16*).

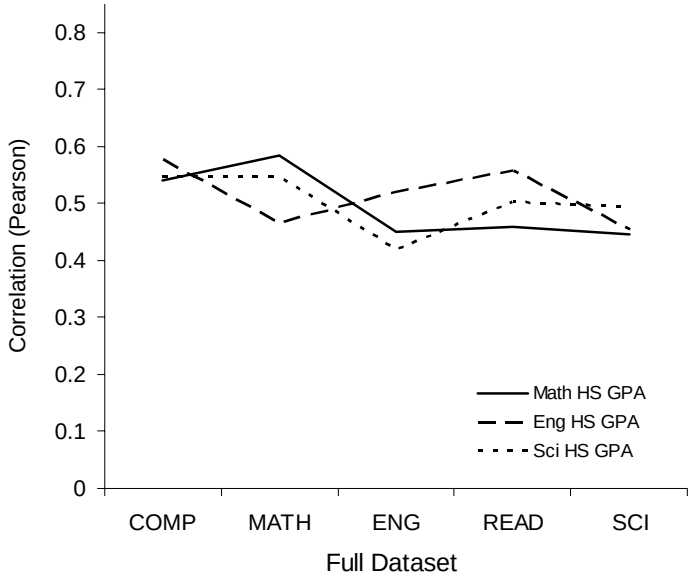


Figure 15: *Correlations of high school subject specific GPA with the ACT subtest, full dataset*

As in Table 14, subject-specific high school GPAs moderately correlated across all of the ACT subtests for the full dataset (*Figure 15*) indicating that subject-specific GPA correlates similarly to grade-level subject-specific grades, and thus may be a more interesting variable to correlate with ACT scores since subject-specific GPA does not include grades from non-core subjects like HSGPA does. Subject-specific GPA correlation with ACT subtest scores varied between 0.4 and 0.6 and followed the pattern described in the tables and figures above in that mathematics GPA correlated higher than English and science with the mathematics ACT subtest. A similar trend was repeated for the other subjects. As with the data presented above for non-similar subjects, mathematics and science GPA correlated the least with the English ACT subtest, while English GPA correlated the least with the mathematics ACT subtest, and English and science GPA correlated similarly and lower than science GPA with the ACT science

subtest (*Figure 15*). Regardless, the correlations of the full dataset do not address the question of a change in correlation overtime, so a comparison of cohorts is required.

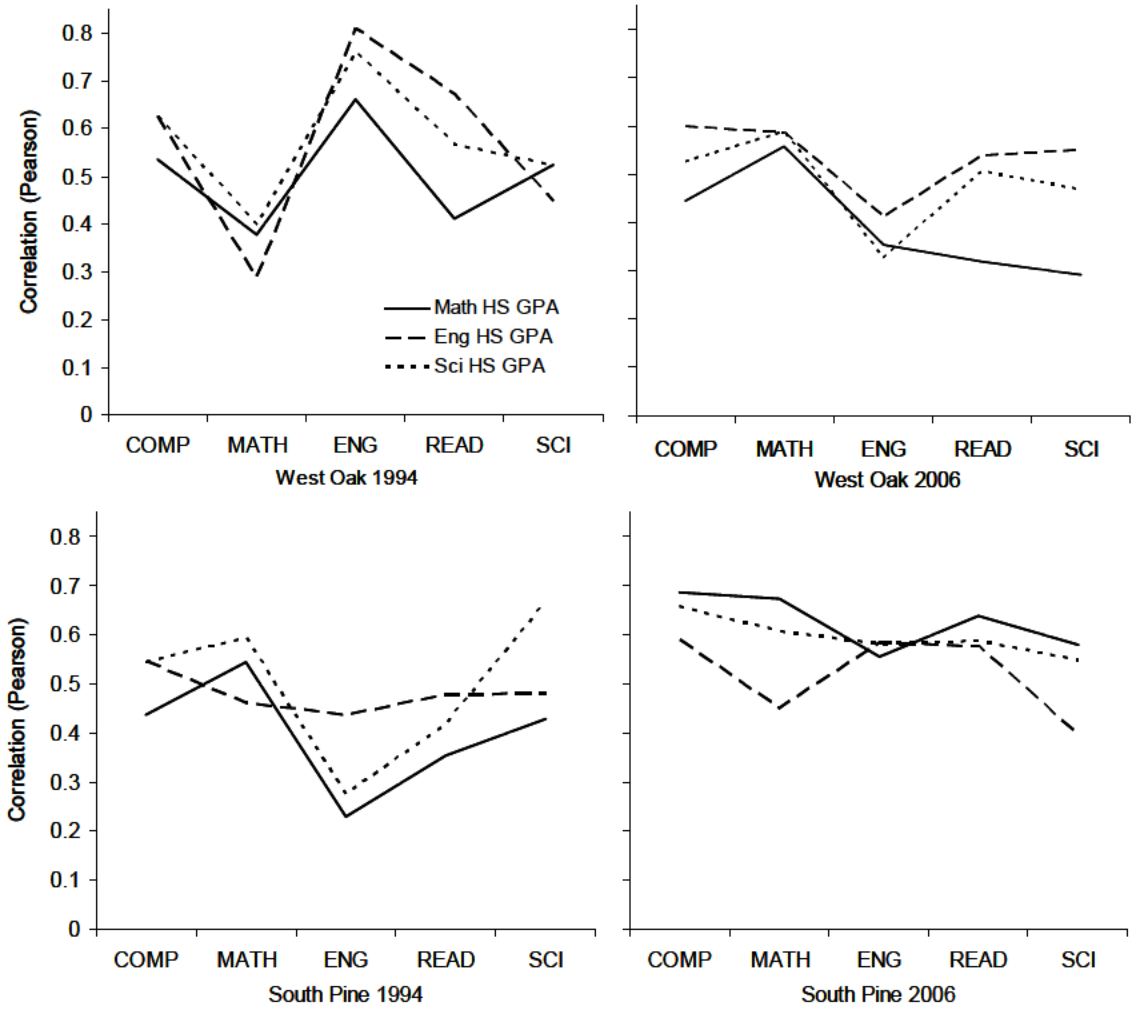


Figure 16: *Correlations of subject-specific high school GPA and ACT between the 1994 and 2006 cohorts for both West Oak and South Pine (Pearson correlations)*

Correlations for high school subject-specific GPA for mathematics, English and science to the ACT subtests were disaggregated by cohort and district (*Figure 16*). As in the data presented above, correlation trends for West Oak indicate that the overall correlation of grades and ACT scores decreased between the 1994 and 2006 cohorts (*Figure 16, top panels*). For West Oak in specific ACT subtests there were striking

differences between the 1994 and 2006 cohorts in the correlation of all three subject-specific GPAs with the mathematics, English and reading ACT subtests. Specifically, the correlation between all three subject-specific GPAs and the ACT English subtest has decreased, from the 1994 cohort in which all three correlations were over 0.6, to the 2006 cohort in which all three correlations were 0.5 (*Figure 16, compare all three lines in the top left panel ENG column with top right panel ENG column*). Also, the correlation between English GPA and the ACT reading subtest has decreased between the two cohorts. Interestingly, the correlation between all three subject-specific GPAs and the ACT mathematics subtest has increased between the 1994 cohort and 2006 cohort, rising from lows under 0.4 to highs over 0.5 (*Figure 16, compare all three lines in the top left panel MATH column with the top right panel MATH column*). However, the overall patterns for West Oak in Figure 16 suggest that the correlations between subject-specific GPA and ACT scores has not substantially increased from the 1994 cohort to the 2006 cohort across multiple subjects and tests. It must be noted however, that confidence intervals for each of the correlations for each of the 1994 cohorts overlaps with the comparison 2006 cohorts (*see Appendix B*). This most likely is due to the small sample sizes of just the students in each cohort who took the ACT, but does indicate that all of the differences between the cohorts are not statistically significantly different.

In contrast to West Oak, the data for South Pine indicates that the correlations between subject-specific GPAs and ACT scores may have increased between the 1994 and 2006 cohorts (*Figure 16, bottom panels*). While the lowest two correlations for South Pine were for the 1994 cohort in mathematics and science GPA correlated to the ACT English subtest (*Figure 16, bottom left panel*), the South Pine 2006 cohort appears to

have higher correlations than the 1994 cohort in the majority of the subject-specific GPAs and ACT subtests (*Figure 16, bottom right panel*). The evidence presented here shows a general but statistically non-significant increase in the correlation of subject-specific GPAs to ACT scores for South Pine but not West Oak.

To further explore these differences in correlations, 10th grade semester 2 grades in mathematics, English, and science were correlated to the ACT subtest scores for both cohorts for both districts using Spearman's Rho correlation (*Figure 17 and Figure 18*). As stated above, 10th grade semester 2 grades are a relevant comparison to ACT scores, since students take the ACT during the following academic year after the 10th grade semester 2 grades were assigned. For West Oak, the correlations across multiple ACT subtests of teacher assigned grades in mathematics, English and science are highly variable and show that generally the correlations for the 2006 cohort are below the correlations for the 1994 cohort (*Figure 17, compare dashed lines for the 1994 cohort to solid lines for the 2006 cohort*). This is most obvious in English and science, in which the correlations between English and science grades with ACT subtests in English and science respectively are lower for the 2006 cohort than they are for the 1994 cohort (*Figure 17, center and bottom panels*). Only the correlation between mathematics and English grades and the ACT mathematics subtest are appreciably higher for the 2006 cohort than the 1994 cohort (*Figure 17, top and center panels*), however for all of these correlation comparisons, confidence intervals overlap and thus are not statistically significantly different (*see Appendix B*).

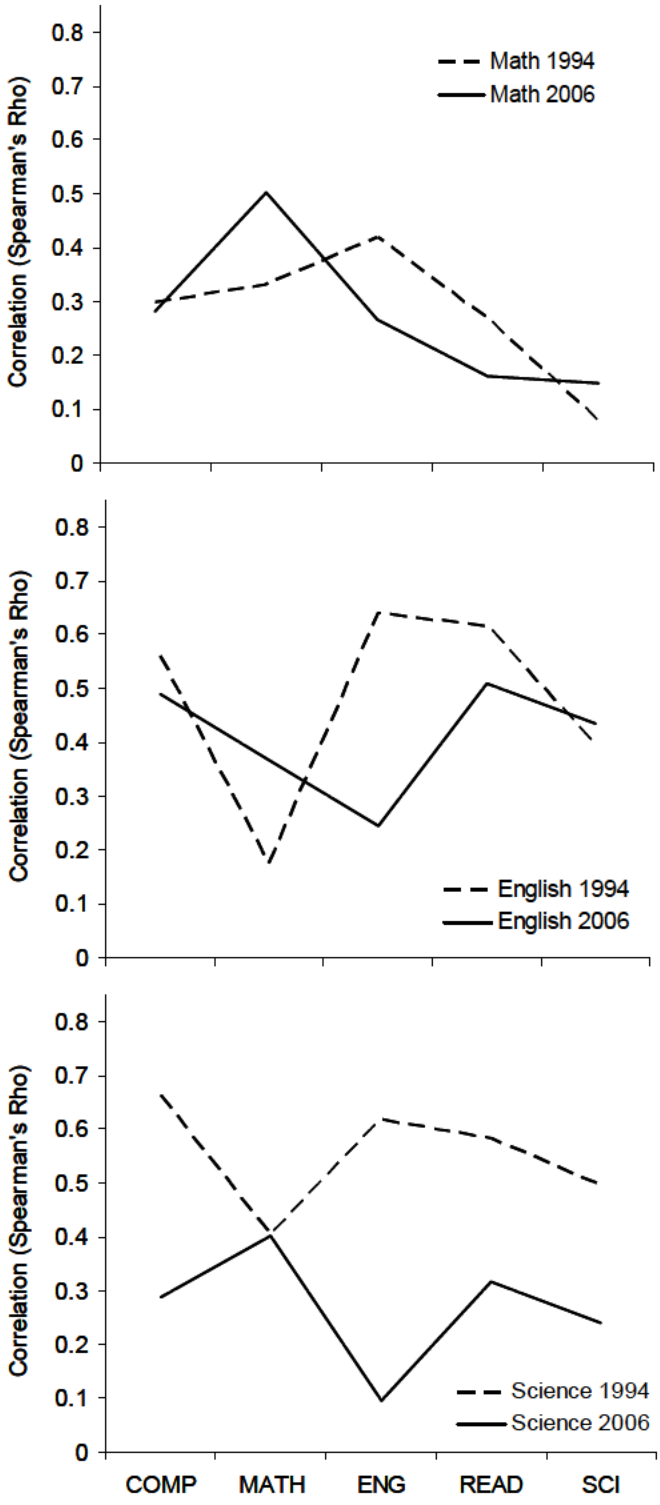


Figure 17: *Correlations of West Oak 1994 and 2006 cohort 10th grade semester 2 subject-specific grades in mathematics, English and Science to ACT subtests (Spearman's Rho)*

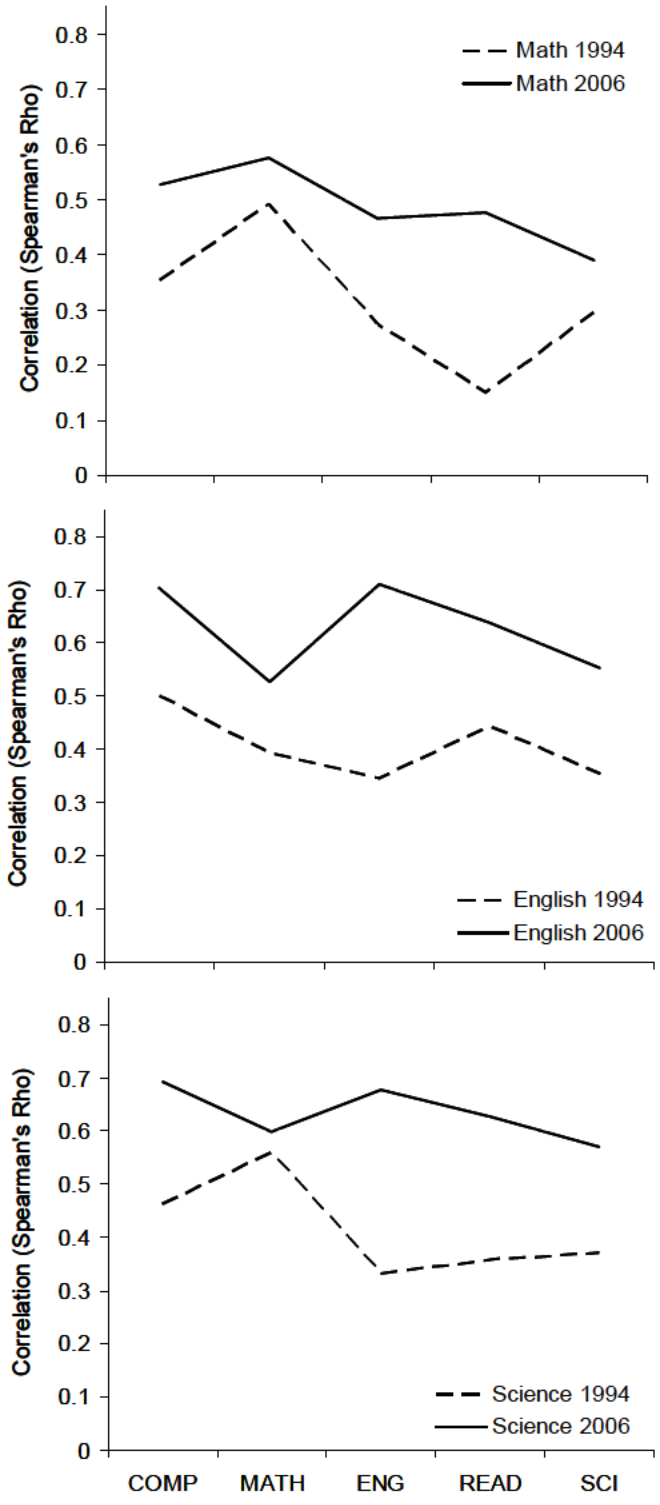


Figure 18: *Correlations of South Pine 1994 and 2006 cohort 10th grade semester 2 subject-specific grades in mathematics, English and Science to ACT subtests (Spearman's Rho)*

Similar to the data presented in Figure 16 above, the correlations for South Pine between 10th grade semester 2 subject-specific grades in mathematics, English and science to the ACT subtests are somewhat higher across subjects for the 2006 cohort in comparison to the 1994 cohort (*Figure 18*). For grades in each subject examined, 2006 cohort correlations to the ACT subtests exceeded the correlations of the 1994 cohort (*Figure 18, compare dashed lines to solid lines*), however, as with the West Oak data, confidence intervals for each correlation comparison overlap indicating that all differences are not statistically significant (*see Appendix B*). This data suggests that while West Oak has not seen an overall increase in the correlation between grades and ACT scores, South Pine has seen a slight but statistically non-significant increase in correlations; however that increase is only for the correlations between core subject grades and the ACT; namely mathematics, English and science.

The research question for this chapter is: to what extent has the correlation between grades and standardized assessments changed from earlier student cohorts to more recent cohorts? Based on data from two districts for two cohorts, each separated by 12 years the evidence is mixed. West Oak has not seen an appreciable increase in the correlation between grades and ACT scores, while the data suggests that South Pine has seen a non-statistically significant increase, at the least for the core subjects of mathematics, English and science. This point must be further critiqued in that the entire burden of the correlations presented in this chapter have rested not on state test scores administered to each student (a better but impossible option due to the test records themselves) but on the ACT scores of a subset of the sample of students from each cohort (those who took the ACT). This issue greatly weakens the finding that the South Pine

data may support the hypothesis. Additionally, the differences in the South Pine correlations between the 1994 and 2006 cohorts are not statistically significant. This may be due to the small sample sizes, but also serves to weaken support for the hypothesis.

However, if one considers this study a pilot study, and the data from West Oak and South Pine merely base-line data, as is suggested in chapter III, then these findings are encouraging and suggest further work. It appears that for South Pine, the correlation between grades and a standardized assessment may have increased over time, providing initial confirming evidence for the first hypothesis proposed in chapter III. Caveats to this finding, as well as a broader discussion and suggestions for future work are discussed further in the final chapter.

CHAPTER VI: GRADE PATTERNING AND PREDICTION

Can grades be used for data driven decision making? This is the primary question addressed in this study. The data presented in chapter V show that grades appear to be converging with one form of standardized assessment for one of the districts in this study suggesting that they have some potential for data driven decision making. The initial data presented suggest that grades may measure more about a student's performance in the schooling system than standardized tests historically have. Specifically, for this sample, there is tentative evidence that grades might be measuring both academic knowledge and success at schooling, argued above as two separate components of teacher assigned subject-specific grades. Hence, rather than being subjective and irrelevant measures (as much of the literature on grading would lead one to believe) grades appear to measure these two variables. This, of course, is an empirical question, one worthy of further examination. This study now turns to another aspect of the study, to demonstrate that grades can be used by school leaders to make decisions that positively impact the lives of students. Toward that purpose, this chapter turns to the next two sets of research questions; to what extent do previous grading patterns predict future grading patterns, and to what extent are grading patterns predictive of qualitative student outcomes, such as on time graduation? These are important questions to consider in relation to data driven decision making since prediction of future performance at a point early in a student's schooling allows for interventions by teachers and administrators, if necessary, and since we know that graduating from high school is a good predictor of a student's life outcomes (Kienzi & Kena, 2006). If grade patterns are useful in predicting on-time graduation or not on time graduation (NOTG) at the earlier stages of schooling, district

and school leaders would gain an additional tool for to help identify students who may need more focused attention by the district. Since these two research questions deal with grade patterns, and the predictive ability of these grade patterns on future student grades and qualitative outcomes such as on-time graduation, these two questions will be considered in tandem as the data is presented.

Not On-Time Graduation (NOTG)

The primary qualitative outcome that this study will focus on is "not on time graduation" (NOTG). For this sample, NOTG is used rather than "dropping out", primary because 1) the term and measurement of "dropout" is currently contested in the literature, and 2) for the data collected, those students who did not have evidence of on-time graduation in their permanent files were assigned to the NOTG category. Because of the issues of identifying NOTG students, it must be assumed that some proportion of the NOTG students are false positives, most likely resulting from the student having a valid transfer to another school district and no record of that transfer existing in the student's files. Despite this issue, the NOTG variable is an indication that the student did not graduate on time with their cohort in either of the two districts. While the false positive issue is a threat to the internal validity of the conclusions of this study because the number of false positives can not be estimated, NOTG is a reasonable designation given that the majority of the students coded NOTG did have records of either non-attendance, refusing to attend the school, incarceration or expulsion. In this way, NOTG, while not a "pure" indication of dropping out, should be considered a reasonable proxy. In addition, as mentioned in the methods, an unknown segment of on time graduators may also be false positives, due to some students transferring to other school districts and their

graduation status becoming unknown. Before discussing student grade patterns and how those patterns may predict NOTG, the qualitative outcomes of NOTG and on-time graduation for the dataset will be first detailed.

Students who graduate from high school and receive a regular diploma, on average, experience better life outcomes in terms of employment, type of job, and salary as well as lower rates of public assistance and incarceration (Dynarski & Gleason, 2002; Jimerson *et al.*, 2000; Kienzi & Kena, 2006; Laird *et al.*, 2006; Lehr *et al.*, 2003). The research literature to date examining student graduation has focused on large-scale estimations of national graduation and dropout rates. For the 2003-2004 school year, the United States Department of Education estimated a national graduation rate of 74.3% (Seastrom *et al.*, 2006), and that data is supported by other studies that have also estimated national average graduation rates above 70% (Greene & Caire, 2001; Greene & Winters, 2005). However, other recent studies have begun to reexamine the methods of national graduation estimation and have reported national average graduation rates below 70% (Swanson, 2004). Applying these broader measures of graduation rates, using the NOTG data in this study to calculate on-time graduation rates for South Pine, rates have increased from 63.1% in 1994 to 87.8% in 2006, while the graduation rate for West Oak in 2006 was 65.7%. The graduation rate for West Oak in 1994 can not be calculated since the 1994 cohort data files had been purged of all students who did not graduate on-time with their cohort, as described above. This high variability over years, as well as between districts is reflective of the national debate on average graduation rates for all districts, and is thus not unexpected. It replicates the more general national averages and extends the findings on graduation rates to the individual district level for this sample.

To delve further into the NOTG data, the percentages of NOTG students disaggregated by IEP status, gender and ethnicity is shown in Table 16.

Table 16: *Descriptive variables and frequencies by district and cohort year for students who did not graduate on-time (NOTG)*

	West Oak	South Pine	
NOTG Descriptive Variables	2006	1994	2006
Percent with IEPs	22.2	16.7	27.3
Gender (%)			
Female	36.1	35.4	45.5
Male	63.9	64.6	54.4
Ethnicity (%)			
European American	28.6	91.7	37.5
Hispanic	42.9	5.6	12.5
African American	28.6	2.8	37.5
Asian	N/A	0	0
Multi-ethnic	0	0	0

Since the West Oak 1994 cohort only includes students who graduated on-time, that cohort is not included in Table 16. For the other three cohorts, the data is striking. Across all three cohorts, of the students who did not graduate on-time (NOTG), males consistently graduate on-time at lower rates than females, as do Hispanics and African Americans graduate at lower rates than European Americans and Asians (*compare Table 16 and Table 10 overall demographic variables*). These findings replicate previous studies and extend the findings to the context of small first-ring suburbs. Previous studies have focused on large urban districts, namely Chicago and Baltimore, and have shown that the students who most frequently do not graduate on-time are males, Hispanics and African Americans (Alexander *et al.*, 2001; Allensworth, 2005; Allensworth & Easton, 2005; Campbell, 2004; Roderick & Camburn, 1999). An examination of the broader U.S. population has shown that, for the U.S. as a whole, on average there is no difference in

on-time graduation rates between females and males, but that on-time graduation rates for Hispanics and African Americans are much lower than for other ethnic groups (Laird *et al.*, 2006), as confirmed in this study (*Table 16*). It should be noted that to conceal the identity of the students and districts, absolute numbers for each category will not be discussed, however, for many of the categories in Tables 16 and 17, the number of students in any one cohort in any one district may be only in the single digits.

Table 17: *Descriptive variables and frequencies by district and cohort year for students who were retained*

		West Oak	South Pine	
Retained Student Descriptive Variables	Overall	2006	1994	2006
NOTG (%)	85.2	81.8	100	84.6
IEPs (%)	25.9	18.2	33.3	30.8
Gender (%)				
Female	33.3	18.2	66.7	38.5
Male	66.7	81.8	33.3	61.5
Ethnicity (%)				
European American	36.8	28.6	33.3	44.4
Hispanic	21.1	14.3	33.3	22.2
African American	42.1	57.1	33.3	33.3
Asian	0	N/A	0	0
Multi-ethnic	0	0	0	0

Interestingly, the literature to date on dropouts and on-time graduation has indicated that student grade retention is a strong predictor of a student not graduating on-time (Jimerson *et al.*, 2002; Jimerson *et al.*, 2005; Laird et al., 2006; Montes & Lehmann, 2004; Roderick & Camburn, 1999; Roderick *et al.*, 2000). For this study, Table 17 presents data on descriptive variables for the students retained in the three cohorts for which NOTG data was available. Students who were retained and were included in this

study were students who began 1st grade at the same time as the rest of their cohort, and were subsequently held back one, or multiple years. Of the students who were retained, 85.2% of them did not graduate, 25.9% of them had IEPs, 33.3% of them were Female, and 66.7% of them were male. If retentions were random and reflective of the overall demographic characteristics of the population, one would expect student retentions disaggregated by ethnic group to reflect the overall population demographics. However, as detailed in Table 17, a disproportionate percentage of African Americans were retained, 29.6%, in relation to the overall representation of African American students in the sample population, 6.1% (*compare Table 10 and Table 17*) and the same trend is true for males. These findings again replicate and extend previous findings in the literature to the context of these two school districts, indicating that student grade retention is a strong predictor of NOTG.

Retaining a student at any grade level is one of the best predictors of dropping out (Laird et al., 2006; Montes & Lehmann, 2004) and thus also NOTG, as shown in Table 17. The literature on risk factors that predict dropping out also include many other variables that have been tested for the ability to assign students as "at-risk" with the purpose of predicting, and ultimately preventing future student dropouts. However, the predictive validity of these risk factors is known to be relatively low (Dynarski & Gleason, 2002; Gleason & Dynarksi, 2002). Some of these risk factors are a single parent home, family on public assistance, sibling drop out, absenteeism, disciplinary problems, or overage for grade-level, among others. However, individual dropout rates for students with each risk factor have all been shown to be below 10% of the students with that risk factor at the middle school level, and below 30% at the high school level (Gleason &

Dynarksi, 2002; Laird et al., 2006; Montes & Lehmann, 2004; Weber, 1989). If many of these factors are combined using multivariate statistics, the percentage of students identified with the multivariate prediction variable who ultimately drop out, rises to 23% at the middle school level, and 42% at the high school level (Gleason & Dynarksi, 2002). Also, failing grades at the high school level have been identified as a major risk factor of student dropout (Allensworth, 2005; Allensworth & Easton, 2005). However, all of these risk factors only accurately identify a subset of the students who ultimately dropout.

These studies are limited in that the vast majority of the studies have only included data on students at the high school level, and to a much lesser extent middle school or earlier. This is problematic. If identification of potential dropouts does not occur until high school, the deleterious impact of these risk factors over the extended period of time before high school is not assessed or included when judging early risk factors. The literature on student's lack of motivation to stay in school indicates that the decision to dropout is not based on a single factor or moment, but rather is the cumulative effect of multiple risk factors, influencing the student over long periods of time within a district (Jimerson *et al.*, 2000). For the districts in this study, as for many districts nation-wide, early student potential dropout identification is critically important so that the district can intervene. However, districts lack a cheap and effective method for early identification of potential dropouts, earlier than high school, which is able to identify a high proportion of potential dropouts. Studies using current risk factors are successful in identifying only 30-40% of students who ultimately drop out. These results suggest that a district would be unable to identify 60-70% or more of the district's potential dropouts, and may be providing dropout prevention services to a population of students who most

likely would not have dropped out (Gleason & Dynarksi, 2002). One assertion of this study is that what is needed is an advance over current risk factors that uses data that already exists in schools, that is rapid and cheap (Gleason & Dynarksi, 2002), and that identifies a higher percentage of potential dropouts at an earlier stage in school than high school. The examination of grades and grade patterns through cluster analysis meets these specifications.

Student patterns of teacher assigned grades are a potentially rich data source which may have the ability to predict NOTG in a student's early schooling career in a district. This statement, combined with the possibility that past student grading patterns might predict future student grading patterns, is the focus of the remainder of the chapter. Additionally, to focus the discussion, and remain centered on the two remaining research questions, which refer to overall dataset patterns rather than on district specific questions, the following cluster analysis of the dataset will include results only on the overall dataset. Examining each district and cohort using the methods detailed below is of interest, but is outside the scope of this study.

Cluster analysis of grades

Hierarchical cluster analysis of the entire K-12 teacher assigned subject-specific grading histories for the full dataset was used to address the two remaining research questions detailed in chapter III: to what extent do past grade histories predict future grading histories, and can past grade patterns predict student qualitative outcomes, such as on-time graduation?

Cluster analysis has been rarely used in education. The statistical method has the potential, however, to help define natural patterns within student and school-level data

that can be informative for data driven decision making. Examining school-wide student data patterns to better address school-wide improvement and student needs has been suggested in the past (Lortie, 1975; Schmoker, 1999), but an empirically driven statistical method to examine such patterns has been lacking for data driven decision making in schools. Cluster analysis can serve this purpose. In short, hierarchical clustering can address the issue of whether past grading histories predict future student grades. Once cluster patterns are defined, analysis of categorical variables for students, such as NOTG, can be compared to the clusters, examining if grade pattern specific clusters of students show a relationship with either on-time graduation or NOTG. This type of analysis, of first clustering and then examining if pattern groups relate with categorical data, was pioneered in the biological taxonomy field (Sneath & Sokal, 1973). It has more recently gained significant popularity in cancer research and molecular pharmacology as an attractive statistical technique for organizing extremely large datasets in which hundreds, if not thousands, of variables are collected on hundreds of patients to help predict patient outcomes and possible intervention strategies (Bowers et al., 2000; Eisen et al., 1998; Kallioniemi, 2002; Lu *et al.*, 2005; van'tVeer et al., 2002; Weinstein et al., 1997). In one of the earliest cancer studies, researchers analyzed if 5000 different genes were turned on or off in 98 different breast cancer tumors from different patients (van'tVeer et al., 2002). They then used hierarchical clustering on this dataset to give organization to the data. Once large patterns of gene expression were defined, the researchers found that specific clusters of gene expression patterns correlated with a poor patient prognosis, indicating possible avenues for therapeutic and diagnostic research, using the information about which genes patterned into specific clusters for either a good or poor prognosis. This

work has recently been extended, classifying previously difficult to classify tumors using cluster analysis of genetic patterns (Lu *et al.*, 2005).

The cluster analysis employed in this study is of a highly similar nature to the cancer studies, but uses students and their grades rather than patients and tumor genes. The aim is to define specific clusters of student grade patterns from the past which predict specific outcomes, such as on-time, or not on time (NOTG) graduation. Specific grade patterns may be predictive of the specific NOTG outcome, indicating a use for grades in predicting NOTG, as well as demonstrating that early grade patterns predict future grade patterns, the two research questions addressed in this chapter.

In addition to these insights offered by cluster analysis, cluster analysis is a descriptive statistical analysis, having fewer of the statistical assumption problems of multiple regression which were detailed in chapter II. In cluster analysis, in comparison with multiple regression using district-level data, the violated assumptions of multicolinarity, variable and case dependence, and nested data are all positives, giving the underlying structure to the data that hierarchical clustering aims to uncover. Also, as discussed in chapter II, many leaders in schools and districts have little interest in generalizing their data to the population mean, the object of inferential statistical analysis such as multiple regression and HLM. However, these decision makers are very interested in descriptive statistics, which are able to reveal actionable analyses of their students in real time. This study aims to show that hierarchical clustering can help leaders in this regard.

The supposition that drives this chapter is that the hierarchical clustering of the entire grading histories of the students in the full dataset should be able to define, at the

minimum, two clusters: those who graduate on time, and those who do not. In addition, a cluster of on-time graduating students should also correspond to students who had taken the ACT, while a NOTG cluster would correspond to students who were retained or who did not take the ACT, described above in chapter V.

Hierarchical clustering of grades

Teacher assigned subject-specific grades for each student K-12 in the full dataset were clustered according to the methods and plotted on an Eisenplot (*Figure 19*). Individual students are on the vertical axis, while subjects per grade level are on the horizontal axis (*Figure 19, center*). Z-scored student grades are represented as a heat map across both axis, with more intense red indicating a higher z-score grade, while a more intense blue indicates a lower z-score grade, with grey indicating the mean and white indicating no data (*Figure 19, center*). Hierarchical clustering patterns are represented in a dendrogram (a cluster tree) with the distance measure indicated at the bottom in standard deviation units (*Figure 19, left*). Student grade pattern clusters were then compared with the categorical variables, NOTG, took ACT, retained, transferred into the district at any time, transferred out of the district at any time to a valid school (as described in the methods), female student, attended West Oak, and was part of the 2006 cohort (*Figure 19, right side, black bars*). Each of these variables is dichotomous, such that a black bar represents the presence of that variable for that student, and no black bar represents the absence. School level is indicated at the top along the horizontal axis. Grade-level increases left to right and subjects follow a repeating pattern by grade-level, from core subjects to non-core subjects (*Figure 19, legend*). Student data rows that contain a series of no data, indicated by a long stretch of white, indicate that there was no

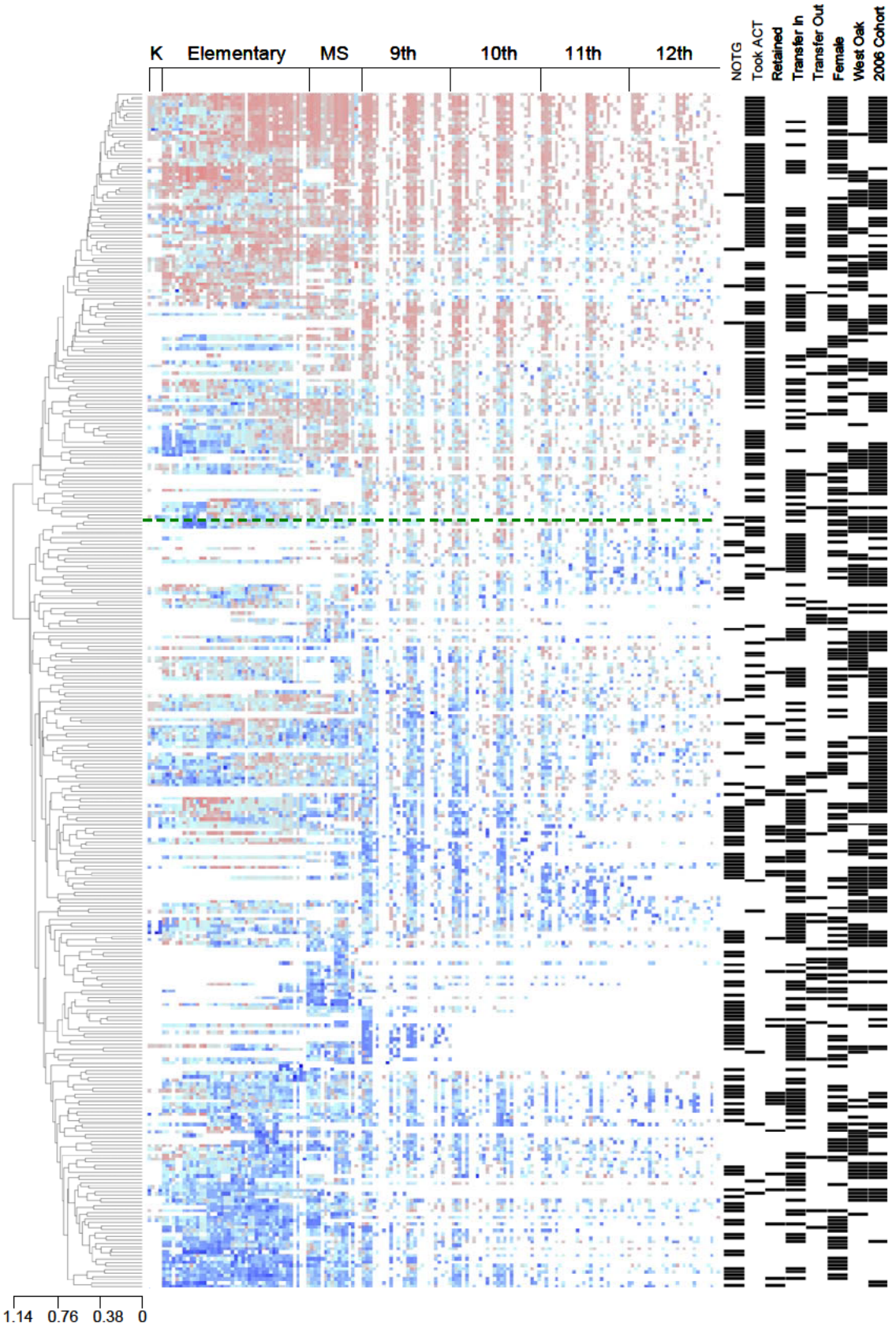
Bowers, A.J. (2007)

data in the file for those grade-levels for that student. If the white row is before middle school, it can be assumed that the student transferred into the district. Date of transfer in can be derived from when the grade color blocks begin. For data rows which contain a stretch of white after middle school, the student either transferred out of the district or was NOTG. The last grade that the student completed can be inferred from the last grade color block that precedes the stretch of white.

Bowers, A.J. (2007)

Figure 19: *Eisenplot of hierarchical clustering of teacher assigned subject-specific grades, full dataset.*

Cluster analysis of student grades (following page) indicates that for the full dataset, student grade patterns cluster into two main clusters, those who graduate on-time, and a high percentage of students who do not graduate on time (NOTG). Each student is aligned along the vertical axis, with subjects by grade-level aligned along the horizontal axis. **This figure is presented in color.** Z-scored student grades are represented by a heat map, with higher grades indicated by an increasing intensity of red, lower grades indicated by increasing intensity of blue, the mean indicated by grey, and white indicates no data (center). Hierarchical clusters are represented by a dendrogram (left), with a scale in standard deviation units for the clusters across the hyperdimensional dataspace in standard deviation units (bottom left). Dichotomous categorical variables are represented by black bars for each of the categorical variables listed (right) as described in the text. The dashed green line through the center heat map indicates the division line between the two major clusters in the full dataset (center). School and grade-level is indicated along the top horizontal axis (center top). Grade level increases left to right, starting with Kindergarten (K), Elementary includes grades 1, 2, 3, 4, 5, and 6, followed by Middle School (MS) including grades 7 and 8, followed by high school and grades 9, 10, 11 and 12. Within each high school grade-level two separate semesters are represented, with semester 1 followed by semester 2. Within each grade-level, subjects are listed in a repeating pattern as follows: K – mathematics, speaking, writing, reading; Elementary - 1^{st}-5^{th} – mathematics, reading, writing, spelling, handwriting, science, social studies; 6^{th} – reading, mathematics, English, science, band, social studies, physical education, art; Middle School - 7^{th} – mathematics, English, science, social studies, band, physical education, health, art; 8^{th} – mathematics, English, science, social studies, band, physical education, study skills, art; high school - 9^{th} – semester 1 – mathematics, English, science, foreign language, social studies, government, economics, band, physical education/health, computers, life skills, family skills, art. Semester 2 repeats 9^{th} semester 1. All other high school grade levels repeat 9^{th} grade subject patterns.



Bowers, A.J. (2007)

The hierarchical cluster analysis of the full dataset indicates that the grading data pattern into two main clusters, those who predominately graduate on-time, and those that have a high percentage of NOTG students (*Figure 19, dendrogram and dashed green line, compare NOTG column above and below the dashed green line*). These two clusters are over one standard deviation from each other in the hyperdimensional grading dataspace (*Figure 19, left bottom*). Students in the top cluster appear to have overall grade patterns that are over the mean for the dataset, as indicated by the majority of red grade data blocks, in comparison to students in the bottom cluster who mostly scored below or at the mean for the dataset in multiple courses, as indicated by grey and blue grade data blocks (*Figure 19, center*). One way to simplify analysis of Figure 19 is for the reader to take a blank sheet of white paper and cover all but the NOTG data column on the right (*Figure 19, right*). With just the one column of NOTG data showing, the difference in the propensity of the bottom cluster to contain NOTG students is striking. It appears that the hierarchical clustering algorithm performed well and was able to distinguish between grading patterns which predict on-time graduation and grading patterns which predict NOTG. If the reader then reveals the Took ACT column, the pattern becomes more interesting. The vast majority of the students in the top cluster (above the dashed green line) took the ACT and graduated on-time, as detailed in chapter V. However, the pattern of students in the bottom cluster who took the ACT appears to be more of a gradient, decreasing as one moves down the bottom cluster as the number of NOTG students increases. These patterns show, that for the full dataset, grades alone, when analyzed using hierarchical clustering, are useful for predicting both NOTG and ACT participation. Additionally, while grades have historically been viewed as subjective

measures of student performance, for this study, it appears that high grades do predict on-time graduation and ACT participation, while *increasingly* low grades predict NOTG and not taking the ACT, indicating that the lower the grades the more likely the student is to not graduate on time and not take the ACT and thus does not appear to have plans to go to college. Students whose grade patterns are more near the mean over their grading histories within the districts appear to graduate on-time at a somewhat lower rate than the top cluster, and take the ACT somewhat more frequently than the rest of the bottom cluster (*Figure 19 bottom, compare the upper quarter which contains clusters of mostly grey patterns, with the clusters lower in the bottom cluster*). Overall there appear to be three main clusters of data, students whose grade patterns are generally above the mean across subjects and grades levels, students whose grade patterns are close to the mean, and students whose grade patterns are below the mean. Interestingly, the cluster analysis shows that students whose grade patterns are at the mean and below the mean cluster together with a higher proportion of NOTG students than the cluster of students with grade patterns consistently above the mean.

If the reader reveals the next column of categorical variables, it can be seen that the data for the "Retained" variable corresponds to the bottom cluster, not taking the ACT, and NOTG, as would be predicted given the discussion of retention above. Revealing the next two columns, "Transfer In" and "Transfer Out," these two variables indicate either if a student transferred into the district at any time, or had a valid request for transcripts from another school district and thus transferred out of the district. Interestingly, there seems to be little correspondence between transfer in or out status, and the upper and lower grade pattern clusters. High mobility has been studied in the past

as an indicator of a student's potential to have low academic performance and to not graduate on time (Demie, 2002; Demie *et al.*, 2005; Montes & Lehmann, 2004; Wells, 2003), however some of this research has been recently criticized in that other variables may explain lower achievement and not graduating on time rather than mobility, such as SES (Machin *et al.*, 2006; Strand & Demie, 2006). For the data presented here, it appears that student transfer in or out of the school districts studied is not an indicator of performance or on-time graduation.

If the reader reveals the next categorical data column, the "Female" variable will be revealed as an indication of how gender relates to the clustering of student grade patterns. For the entire dataset, it does not appear that females clustered more or less in either the upper or lower clusters. However, if the NOTG and Female columns are compared, it can be seen that for the students in the bottom cluster who were NOTG, many were also not female, indicating that males did not graduate on time more so than females, as was discussed above.

The last two categorical data columns are "West Oak" and "2006 cohort". Through comparing the black bars in these two columns, one can place each student grade pattern into the four cohorts in the dataset (*Figure 19, right-most two columns of right-hand panel*); West Oak 1994 (bar, no bar), West Oak 2006 (bar, bar), South Pine 1994 (no bar, no bar), South Pine 2006 (no bar, bar). Interestingly, while the overall dataset clusters into two main clusters that appear to correspond to on-time graduation versus NOTG, specific cohorts of students from one of the two districts sub-cluster within both the top and bottom clusters as evidenced by a non-random clustered pattern

in the black bars for the last two columns. This may suggest specific sub-cluster district teacher, curriculum or grading policy effects.

Therefore, to answer one of the research questions of this chapter of do student grade patterns predict a qualitative outcome such as NOTG, the answer for this dataset is yes. Additionally, if the question is extended to ask if this method is an advance over past methods of identifying students as "at-risk" of NOTG as described above, the cluster analysis shows that while only 4% of the students in the top cluster were NOTG (1 of every 25), 42% of the students in the bottom cluster were NOTG (1 in every 2.4). Thus, when considering the dataset in total, cluster analysis is as efficient a predictor as the high school multivariate regression methods cited above by Gleason and Dynarski (2002). This finding will be further discussed below, and suggests that this method proposed in this study may be considered superior over past at-risk variables for multiple reasons.

In addition to these longitudinal grading cluster patterns, the two overall clusters also indicate a conclusion about the types of courses the students in the top and bottom clusters were enrolled in. The horizontal axis can be considered to be clustered in both the time and the subject dimensions of the data, in that each grade is listed sequentially left to right (clustered by increasing time), and each subject is listed in a repeating order within each grade with the core subjects of mathematics, English, science, foreign language, and social studies listed first reading left to right before the non-core subjects of band, physical education, life skills, and art, among others. Noting the subject enrollment patterns from chapter V, this is a logical ordering of core and non-core courses. Hence, the horizontal axis can be considered to be clustered according to the algorithm of increasing years of schooling and core versus non-core subjects. This

ordering of the horizontal axis thus places the core course subject columns in Figure 19 to the left-hand side of each grade-level. Also, since each high school grade level contains two semesters of data, a second set of core course subject columns for the second semester is in the center of each grade-level column set. In this way, the overall pattern of course enrollment for clusters of students can be determined for the entire dataset.

Interestingly, this pattern in Figure 19 suggests that students in the upper cluster receive high grades, graduate on-time, take the ACT, and additionally, take core subjects through 11th grade and to some extent into 12th grade. This is evidenced by the more solid columns of red blocks in Figure 19 at the left side and middle of each grade-level, indicating that these students took many core subjects, and received a high grade in them. For the lower cluster, the difference in the pattern of classes the students took at the high school level is striking, especially for the lowest quarter of the lower cluster. While the students in the lower cluster appear to have taken core subjects in the 9th and 10th grade, at 11th grade the pattern diverges from the upper cluster, and it can be seen that the lower cluster students take a much wider variety of subjects. Not surprisingly, a gradient that parallels participation in the ACT appears to be at work in the lower cluster, in that as one proceeds down from the dashed green line, students appear to have been enrolled in fewer core courses in 12th grade, then 11th grade, then 10th grade nearest the bottom of the lower cluster. For the dataset, this result may indicate that students who were receiving low grades and who had a higher probability of NOTG than the upper cluster were taking courses in fewer core subjects at the high school level, especially by 11th and 12th grade.

The other research question of this chapter asks if past grading patterns predict future grading patterns, the Hargris Hypothesis. The cluster analysis provides evidence to

Bowers, A.J. (2007)

support this hypothesis as well. Overall, the answer to the question is yes, but with some caveats. From the data presented in Figure 19, students who are assigned high grades in early elementary school (grades K, 1, 2, and 3) generally receive high grades throughout their career for all four cohorts (*Figure 19, upper cluster*). The same appears to hold true generally for students whose grades early on are at the mean or below the mean, in that those students who receive low grades early appear to continue to receive low grades throughout their schooling career. Thus, this study provides initial empirical evidence to support the Hargris hypothesis discussed in chapters II and III that early high grades, in general, appear to launch a student into a cycle of motivation and achievement, while early low grades appear to lock a student into a continual cycle of low grades and achievement (Hargris, 1990).

Examining specific sub-clusters as illustrative of the overall upper and lower cluster patterns in Figure 19 is useful to help understand these two overall patterns. These two patterns are termed "high-high" and "low-low", indicating high grades in early elementary and in high school with on-time graduation by 12th grade (*Figure 20*), or low grades in early elementary and low grades by high school with a large proportion of students NOTG (*Figure 21*). The high-high cluster pattern, here represented by a sub-cluster of 42 students from the upper cluster in Figure 19, indicates that for this dataset, students who are awarded grades at and above the mean in early elementary school, generally go on to earn high grades across all subjects throughout elementary, into middle school, and throughout high school, with eventual on-time graduation (*Figure 20*). Conversely, the low-low cluster pattern, here represented by a sub-cluster of 32 students from the lower cluster in Figure 19, indicates that for this dataset, students who are

awarded grades at and below the mean in early elementary school, generally go on to earn low grades across all subjects throughout elementary, into middle school and throughout high school, with a high proportion NOTG (*Figure 21*). Thus, the data for this study provides initial evidence which supports the Hargris hypothsis.

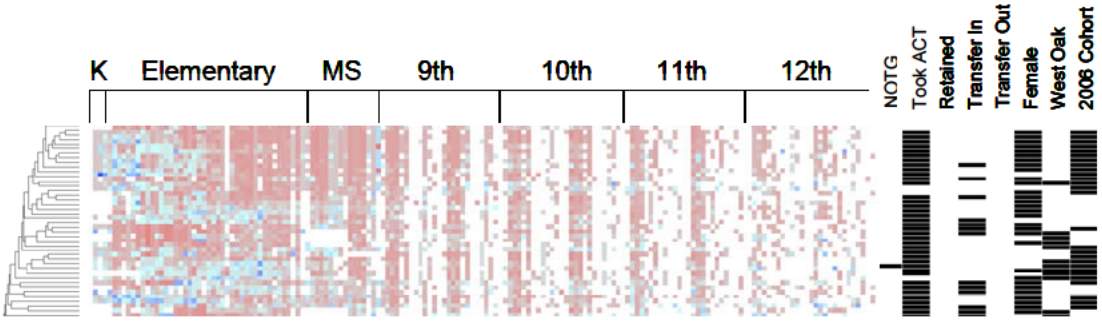


Figure 20: *High-high student grade pattern sub-cluster, K-high school*

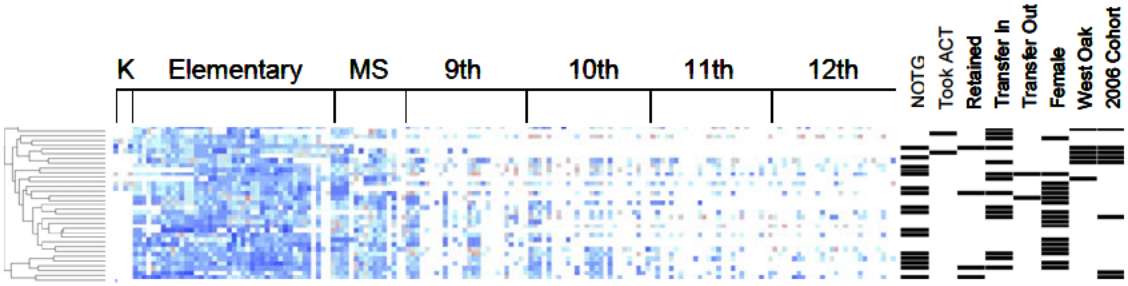


Figure 21: *Low-low student grade pattern sub-cluster, K-high school*

However, the Hargris hypothesis may be only a general pattern observed for the upper and lower clusters in Figure 19. The type of student for whom research has historically been lacking is both the student who starts with high grades in early elementary school but eventually receives low grades (a "high-low" pattern) (*Figure 22*), and the student who starts with low grades in early elementary school but eventually receives high grades (a "low-high" pattern) (*Figure 23*). It appears that clusters of students of these types do exist in this dataset. Further examiniation looks at the relation

of each cluster of students to NOTG. To address these issues, individual sub-clusters from Figure 19 are examined.

First, the low-high pattern will be considered. In the upper cluster of Figure 19, a sub-cluster of 17 students can be identified in which students had a low-high pattern, low early elementary grades in grades K, 1, 2, and 3, that gradually rose to the mean in grades 4 and 5, and exceeded the mean in grade 6 and in middle school (*Figure 22*). As shown in Figure 22, while the overall student grade patterns for the 17 students in the sub-cluster leveled off generally near the mean in high school across the cluster, all of the students graduated on time and most of them took the ACT. Interestingly, the majority of this cluster of students attended South Pine, although they belonged to both the 1994 and 2006 cohorts. Thus, the low-high grade pattern does exist for a subset of the students studied. While low early grades do appear to generally follow the Hargris hypothesis for the full dataset of locking students into a pattern of low grades and a higher probability of NOTG, for the students in the sub-cluster in Figure 22, something happened for them at about the time they attended 2nd and 3rd grades. Some common experience might have helped them improve so that they earned mean grades across subjects by 4th grade, and then continued to improve to grades above the mean in middle school, leveling off somewhat above the mean in high school, and ultimately graduating on time.

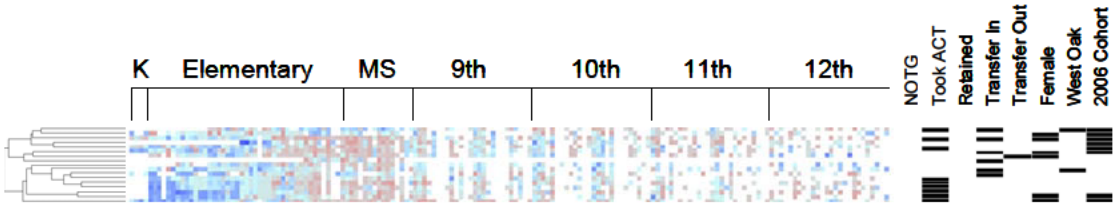


Figure 22: *Low-high student grade pattern subcluster, K-high school*

Similarly, an example of the high-low sub-cluster student grade pattern also exists in the dataset, in which students from the lower cluster in Figure 19 were at or just above the mean grades in multiple subjects at the elementary level, but then their grades drop to below the mean in middle school and high school (*Figure 23*). This pattern appears to be associated with high rates of NOTG and lower rates of ACT participation for all 37 students in the sub-cluster. Thus, it appears that for this sub-cluster of students who achieved at or above the mean in early elementary school, these student's grades eventually begin to fall in late elementary and then throughout middle and into high school, with a high propensity to not graduate on time, nor take the ACT (*Figure 23*).

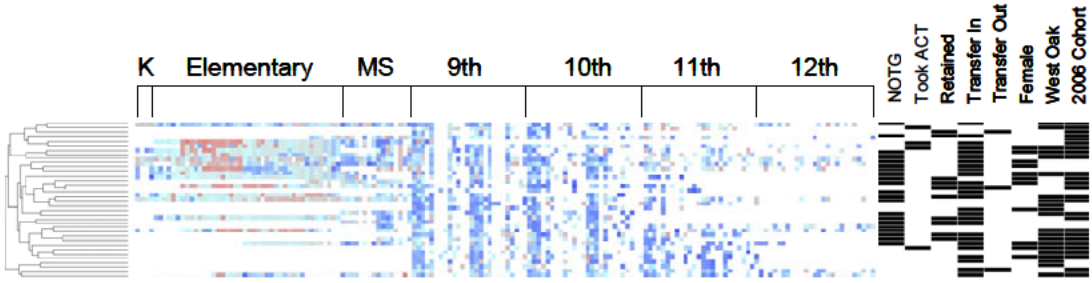


Figure 23: *High-low student grade pattern subcluster, K-high school*

These cluster patterns reflect the z-scored grades of students across each subject for the full dataset. It is of value to examine the actual grades within the patterns of each major cluster identified. This is especially important when considering the variables that predict "at-risk" of not graduating on time. Past research has relied heavily on a student receiving a failing grade in one or multiple subjects (Alexander et al., 2001; Allensworth, 2005; Allensworth & Easton, 2005; Gleason & Dynarksi, 2002; Montes & Lehmann, 2004). A failing grade is predictive of not graduating on-time; however failing grades are not usually given until middle or high school. Most of the studies on failing grades are

only able to analyze high school data; as discussed above, identification of students as "at-risk" at the high school level may be many years after a point when intervention is optimal. If the student was identified as having trouble at a point early in schooling, intervention might have helped that student join a cluster such as the low-high cluster, rather than the high-low or low-low clusters.

For the full dataset, the first failing grade occurs for one student at the 6th grade level, and then the frequency of failing grades slowly rises into the high school years. Remember that the clusters presented above are based on z-scored grades K-12, so year-to-year differences in grading scale are normalized. In fact, a failing grade in high school is a similar in z-scores to a B- or a C+ at the early elementary level. Both an F in high school and a B- in early elementary school are at the bottom of the relative scales at each grade level. The elementary grade however would also indicate satisfactory work even though the student receiving the grade might be at the bottom of the class in terms of performance. An examination of actual grades received across students' school histories help explore these possibilities.

To examine actual grade patterns for students in the upper cluster and lower cluster in Figure 19, the mean non-cumulative GPA for each grade-level for the full dataset were plotted K-12, with high school grades by semester 1 followed by semester 2 (*Figure 24*). Figure 24 shows that the mean GPA of upper and lower cluster students did not converge at any time across all 17 timepoints. Additionally, the trends begin to diverge at and after the 3rd grade-level, with the upper cluster maintaining over a 3.0 GPA (a "B" letter-grade), while the lower cluster declined throughout elementary and middle school, leveling off at a 2.0 (a "C" letter-grade) at the high school level. These grading

trends are especially significant considering that 42% of the lower cluster students were NOTG, while only 4% of the upper cluster students were NOTG.

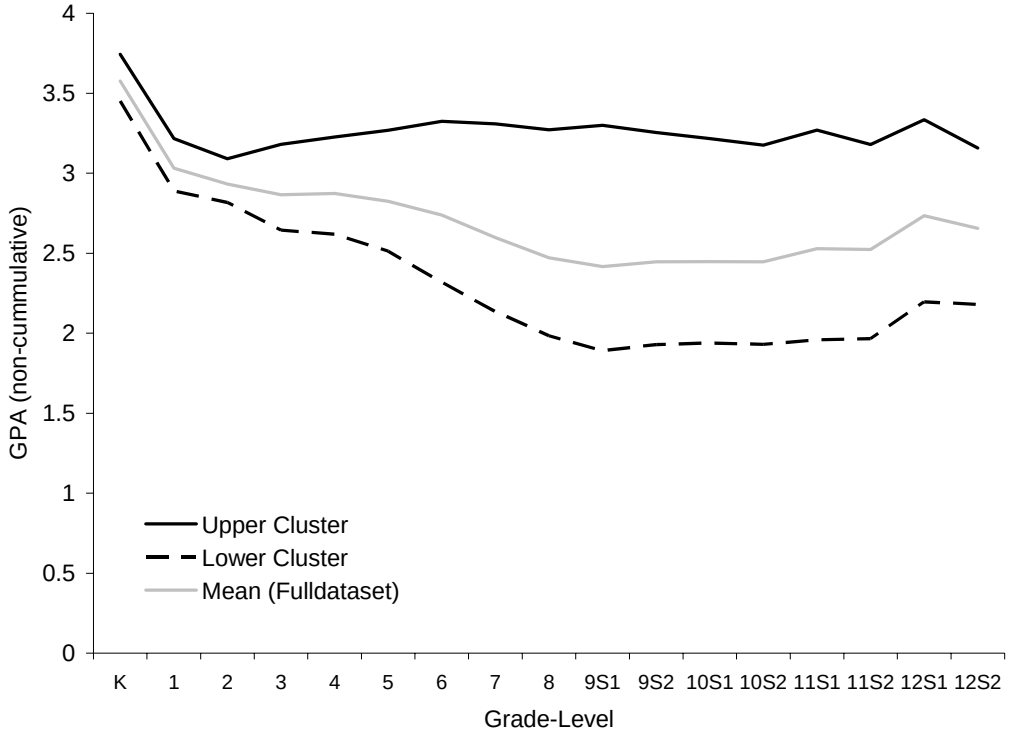


Figure 24: *Mean non-cumulative GPA trends for the upper and lower clusters, K-12*

A similar plot can also be constructed which compares the four above example sub-clusters of student non-cumulative GPA, high-high, high-low, low-high, and low-low (*Figure 25*). While the high-high and low-low clusters in Figure 25 correspond to the trends seen in Figure 24, the high-low and low-high mean cluster GPAs show an interesting trend that has rarely been discussed in the education literature, namely students who start with relatively high grades but then are awarded lower grades over time (high-low), and even less frequently studied, students who start low but then are awarded high grades over time (low-high).

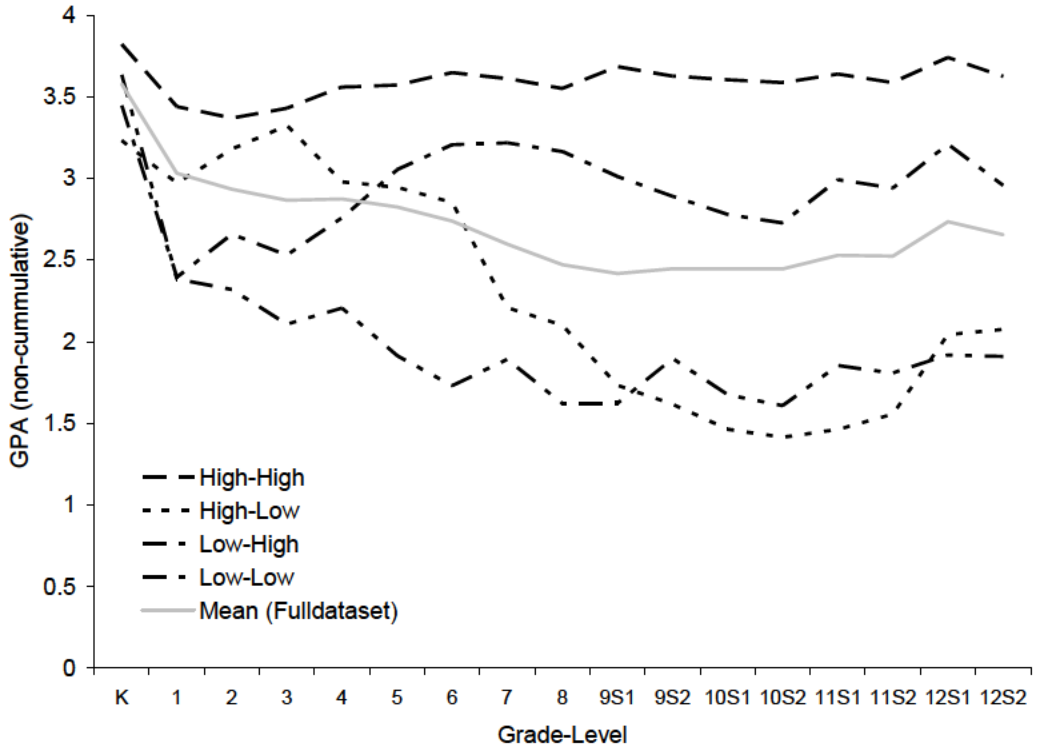


Figure 25: *Mean non-cumulative GPA trends for clusters high-high, high-low, low-high and low-low, K-12*

The overall trends of the four clusters suggest interesting patterns for these four clusters plotted in Figure 25. First, recall that both the high-high and low-high clusters are members of the upper cluster in Figure 19, which corresponds to on-time graduation, while the high-low and low-low clusters are members of the lower cluster in Figure 19, which corresponds to much higher rates of NOTG. Both Figure 24 and Figure 25 indicate that for this dataset, students whose non-cumulative GPAs are below the mean have a much higher chance of NOTG than students whose grades are above the mean (*Figures 21 & 22, compare above and below the grey line*). Second, the low-high pattern is striking, in that the low-high student grades track similarly to the low-low student grades in first grade, but then diverge in second grade and continue to rise, passing the mean,

and the downward trend of the high-low cluster in the 4th grade. The high-low cluster grades start just below the high-high students, but in 4th grade begin to decline, passing the upward trending low-high cluster students and falling below the mean by 6th grade, and trending similarly to the low-low student cluster by high school. The high-high students appear to have been awarded high grades throughout their career, while the low-low students start with some of the lowest grades in the scale in early elementary and continue to receive low grades throughout their career (*Figure 25*). Again, the grading data for the high-high and low-low clusters supports the Hargris hypothesis that past grading patterns predict future grading patterns. However, the high-low and low-high clusters, while subsets of the overall upper and lower cluster patterns (*Figure 19 and Figure 24*), contradict the Hargris hypothesis.

For these two clusters, high-low and low-high, two inflection points in time appear to be evident, 2nd grade for the low-high cluster, and 4th grade for the high-low cluster (*Figure 25*). For the low-high cluster, kindergarten and 1st grade grades are nearly identical to the low-low cluster. If students were identified as at-risk of NOTG on K-1 grades alone, then the students in the low-high cluster would be mis-identified. In 2nd and 3rd grade the low-high student cluster is still well below the mean, however their grades are trending higher. Interestingly, in 4th grade the low-high students surpass the mean and cross the downward trend of the high-low students, and then continue to rise throughout middle school, decline somewhat in high school, and ultimately graduate on time with the high-high students.

For the high-low students, in 1st through 3rd grades these students appear to be on-track with the high-high students, and so if identified as at-risk through early elementary

grades these students would not be identified as lower cluster students. At 4th grade, the high-low student's grades begin to decline, fall below the rising low-high student cluster, then fall dramatically as the students enter middle school.

The central question for these two clusters is: what was the difference in the experiences of these two groups of children? Obviously, because these children's grades patterned similarly to each other in Figure 19, something similar may have occurred for the children within each of the clusters, and the timepoints are indicated in Figure 25, 2nd grade for the low-high cluster and 4th grade for the high-low cluster. Additionally, both of these clusters correspond to the low-low cluster. It appears that something changed for the low-high students in 2nd grade that may have caused them to diverge from the low-low cluster, while something may have happened for the high-low cluster in the 4th grade so that the student grade patterns ultimately join the low-low cluster by high school. Figures 19 and 20 indicate that students from all four cohorts are within both the high-low and low-high clusters, both 1994 and 2006 and West Oak and South Pine, suggesting that what occurred most likely was not due to a cohort effect. Was the similarity in grade patterns due to student aptitude or teacher assistance? What can school leaders and teachers do to help more students join the low-high cluster rather than stay in the low-low cluster? Can the GPA decline of the high-low cluster be prevented? These issues will be taken up in chapter VII.

The analyses reported in this study support Hargris's hypothesis (1990). Generally for this dataset, students who received high grades early in elementary school continued to receive high grades, and students who received low grades early in elementary school continue in a cycle of low grading throughout their career in the school

districts. However, these trends appear to be only general trends, since the data presented above suggest that for a subset of the data, some students who start low do attain high grades at the higher grade levels, and these students appear to graduate on time. Another subset of students start high in the grades they receive, but then receive low grades as they progress through the system. These students appear to not graduate on time as often. This is the first time that student grade patterns have been examined empirically for entire cohorts of students in this way, and the first time that the high-low and low-high patterns of student grades have been explicated for such a dataset. As with the questions posed in the previous paragraph, the implications of the existence of these patterns will be discussed below in chapter VII.

The data presented in this chapter thus far suggests that grade patterns can be used to identify students "at-risk" of NOTG. The methods presented here surpass previously used identification methods in multiple ways, including positively identifying 42% of the students who did not graduate on time using currently existing data. The grade trends appear to stabilize by high school. The data presented in Figure 24 and 22 suggests that the hierarchical cluster analysis may produce similar results as those in Figure 19, even without the high school data included. It appears from the data presented above that the grading patterns of the upper and lower clusters vary in elementary and middle school, especially for the clusters of students who do not conform to the Hargris hypothesis; the low-high and high-low students. But, by the end of middle school this variance subsides and student grade trends appear to remain relatively stable throughout high school.

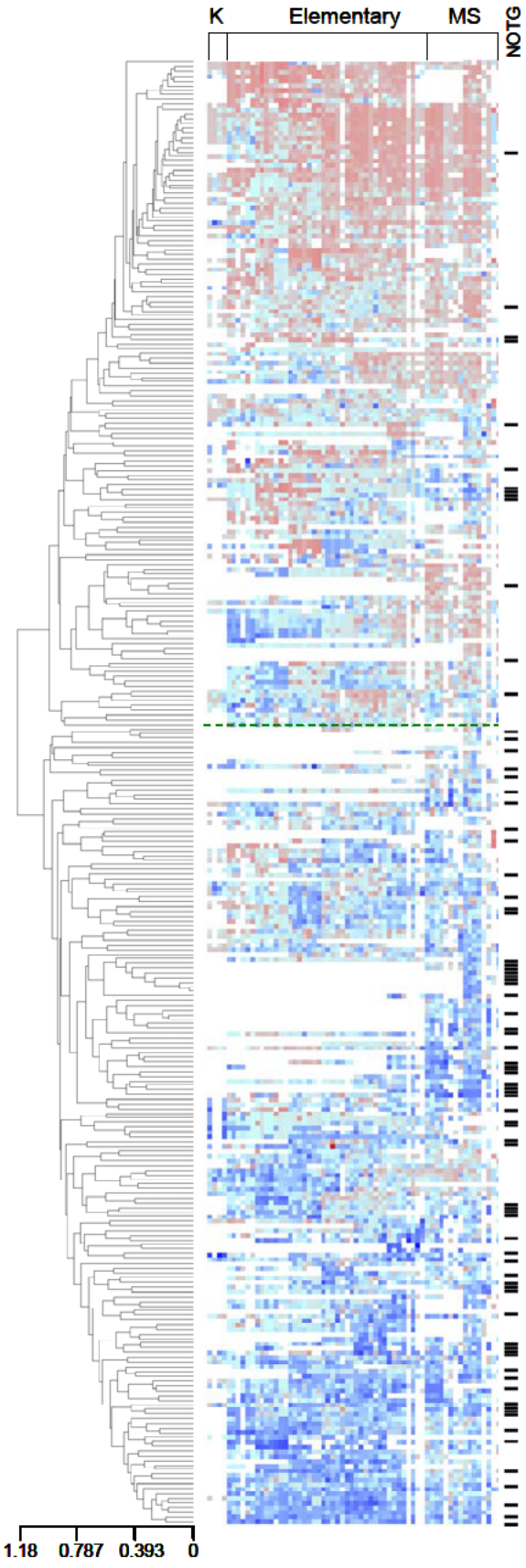
With this knowledge, this study can turn to the question of how efficient a predictor are student grade patterns before high school. This is an important question for

two main reasons. First, as discussed above, early identification of at-risk of NOTG is considered desirable by educational leaders engaged in data driven decision making so that additional assistance can be directed to students who may not graduate on time. Second, the best at-risk prediction variables from the literature are mostly at the high school level, including failing grades. In contrast, in the literature, middle school prediction variables of students at-risk of NOTG overall are considered to be much less accurate than at the high school, with the best methods accurately predicting only about 23% of the students who eventually do not graduate on time (Gleason & Dynarksi, 2002). Thus, a method that identifies students before high school with accuracy higher than 20% would be of value.

To explore these issues, the high school grades were removed from the full dataset, creating a K-8 dataset containing the z-scored subject-specific grades for each student K-8. This K-8 dataset was reclustered using the same hierarchical clustering methods for Figure 19, according to the methods (*Figure 26*). Similar to the clustering of the full dataset above, the K-8 clustering identified two main clusters; students whose K-8 grades were generally high across subjects and who eventually graduate on-time (*Figure 26, center panel upper cluster above the dashed green line*) and students whose K-8 grades were generally low across subjects and grade-levels and who had a higher frequency of eventual NOTG (*Figure 26, center panel lower cluster below the dashed green line*). The cluster dendrogram shows that the data is categorized into these two main clusters (*Figure 26, left panel*), and that students in the lower cluster were more frequently NOTG than the upper cluster (*Figure 26, right column, black bars indicate NOTG*).

Figure 26: *Eisenplot of hierarchical clustering of teacher assigned subject-specific grades, K-8 dataset*

Cluster analysis of student grades (following page) indicates that for the K-8 dataset, 311 student grade patterns cluster into two main clusters, those who eventually graduate on-time, and a high percentage of students who do not graduate on time (NOTG). Each student is aligned along the vertical axis, with subjects by grade-level aligned along the horizontal axis. **This figure is presented in color.** Z-scored student grades are represented by a heat map, with higher grades indicated by an increasing intensity of red, lower grades indicated by increasing intensity of blue, the mean indicated by grey, and white indicates no data (center). Hierarchical clusters are represented by a dendrogram (left), with a scale in standard deviation units for the clusters across the hyperdimensional dataspace (bottom left). The dichotomous categorical variables of NOTG is represented by black bars (right). The dashed green line through the center heat map indicates the division line between the two major clusters in the K-8 dataset (center). School and grade-level is indicated along the top horizontal axis (center top). Grade level increases left to right, starting with Kindergarten (K), then Elementary includes grades 1, 2, 3, 4, 5, and 6, followed by Middle School (MS) including grades 7 and 8. Within each grade-level, subjects are listed in a repeating pattern as follows: K – mathematics, speaking, writing, reading; Elementary - 1st-5th – mathematics, reading, writing, spelling, handwriting, science, social studies; 6th – reading, mathematics, English, science, band, social studies, physical education, art; Middle School - 7th – mathematics, English, science, social studies, band, physical education, health, art; 8th – mathematics, English, science, social studies, band, physical education, study skills, art.


K
Elementary
MS
NOTG
1.18
0.787
0.393
0

In contrast to the differences between past NOTG at-risk prediction methods using high school versus middle school data, cluster analysis of K-8 grades (*Figure 26*) is almost as accurate as cluster analysis of K-12 grades (*Figure 19*) in predicting NOTG. Specifically, as detailed above, past at-risk prediction methods using regression analysis are able to predict at best 42% of the students who would have not graduated on-time using high school data (defined as students dropping out), but using middle school data only 23% of the students who would have not graduated on-time by the end of high school were identified (Gleason & Dynarksi, 2002). However, as previously discussed, early and more accurate identification of students at risk of NOTG is desirable. The data for this study show that by utilizing K-12 grade data, cluster analysis accurately predicts 42% of the NOTG students (*Figure 19*), similar to the literature. Figure 26 extends this level of prediction to the K-8 dataset, in which clusters are again identified, and Table 18 further extends these findings by comparing the findings of Gleason and Dynarski (2002) using current at-risk prediction methods with just high school or middle school data with the cluster methods presented here. Hierarchical clustering of K-8 data shows that 10% of the students in the upper cluster were eventually NOTG (1in 10) (*Figure 26*) and 40% of the students in the lower cluster NOTG (1 in 2.5) (*Table 18*), using only K-8 data. Thus, 1 in 2.5 students whose grades pattern similarly to the students in the lower cluster in Figure 26 are NOTG. This is an advance over past at-risk predictors (*Table 18*), identifying a method which would allow school leaders to better predict students at risk of NOTG *before* they enter high school, by which time the above data shows that the students are relatively stable in their performance and outcomes.

To further explore how early cluster analysis is able to predict NOTG with accuracy that exceeds current methods in the literature, additional cluster analyses were performed using a K-6 and a K-1 dataset, reclustering the data for each smaller dataset (*Appendix C*), and the accuracy of the prediction of the upper and lower clusters was assessed for all four clustered sets of data and compared to the previous findings of Gleason and Dynarski (2002) (*Table 18*).

Table 18: *Cluster prediction accuracy from grades of NOTG by dataset*

Dataset	*Gleason & Dynarski (2002)*	*Lower Cluster*
K-12		
% NOTG	42%	42%
K-8		
% NOTG	23%	40%
K-6		
% NOTG	--	30%
K-1		
% NOTG	--	27%

Cluster analysis of student grades identified two main clusters for all four grading datasets, K-12, K-8, K-6, and K-1, in which the upper cluster corresponded to higher grades and a lower rate of NOTG and the lower cluster corresponded to lower grades and a higher rate of NOTG (*Table 18*). The most accurate prediction of NOTG was the cluster analysis of the K-12 and K-8 datasets. Interestingly, the cluster analysis of the K-6 and K-1 dataset accurately identified over 26% of the students who eventually did not graduate on time. This method, using just kindergarten and grade 1 teacher assigned subject specific grades, exceeds the accuracy of the best at-risk prediction methods described in the literature at the middle school level. This is a significant finding and will be discussed further in chapter VII.

CHAPTER VII: DISCUSSION

The data for this study support five main findings when considering grades as potentially useful for data driven decision making by school and district leaders. 1) Tentative findings were detailed that suggest that grades may be an assessment of both academic knowledge and a success at school factor. 2) Grades and standardized assessments may be converging over time, a finding only partially supported in one of the two districts studied. 3) Past student grade patterns are useful in predicting future student grade patterns, partially supporting the Hargris hypothesis. 4) The Hargris hypothesis does not hold true for all students. One cluster of students who receive high grades in early elementary and middle school, earn increasingly lower grades, and are at risk of NOTG. A different cluster of students receive low grades in early elementary school, but seemingly overcame this early deficit to exhibit rising grades throughout later elementary school and middle school, ultimately graduating on time. 5) Student K-12 longitudinal grade patterning using cluster analysis is as good, or better, a predictor of students at-risk of not graduating on time with their cohort as current at-risk predictors from the literature. Overall, these results show that teacher assigned subject-specific grades, rather than being subjective and unreliable measures of student performance are useful for day to day decisions made by teachers and administrators. These grades should be used, not printed on report cards and then locked away in school basements and forgotten. This study shows that grades are useful as assessments of student school performance and are useful predictors of future grading patterns and on-time graduation. Because of these findings, it is argued here that grades can be used for data driven decision making by

school leaders; informing parents, teachers, principals and central office staff of potential future student grades and on-time graduation.

While the findings of the study are promising, there are multiple issues with the validity and generalizability that require discussion. First and most significant is the biased and intact nature of the student samples. Students were not selected randomly from a large population; rather, two small first-ring suburb districts were selected as a sample of convenience, and two cohorts within those districts (the graduating classes of 1994 and 2006) were selected based on the data available in the student's permanent record folders. The conjecture for this study is that when studying small intact samples, similar to the real-world data analysis performed daily by principals and district administrators in schools across the nation, it is advantageous to include every student for which data exists in a school for the cohorts examined. This eliminates internal validity issues due to sample bias, and as detailed previously above, rather than estimating the population means through inferential statistics such as linear regression, the actual population means for each cohort are known. Thus, it is argued here, that school leaders should be encouraged to include every student in their district in data analysis, rather than choose a sample. This could have very interesting implications, especially for large districts in which student data is warehoused electronically for thousands of children. Access to data on the entire population of interest increases the statistical power of significance tests and allows principals and administrators the ability to understand the data for their students. In this way they don't have to generalize to the mean for all students in the nation in order to say something important about the students in their schools. It must be acknowledged that as with any statistic, generalizability beyond the

sample is problematic. But generalizability is not an issue when principals and district administrators are concerned with their entire sample of students, rather than students outside of their districts. Conversely, in generalizing this study to the broader context of K-12 education across the United States, this issue of small and intact sample size should be taken into account.

For the four cohorts detailed here, the findings of this study are applicable to and inform the two districts of what occurred with the correlation and patterns of grades over time with these four cohorts. But do the findings of this study have any relation to other schools and districts? This is the classic question for all research. This study should be considered a pilot study, with initial but as of yet uncorroborated and unreplicated findings. The findings of this study should inform other school contexts due to three major factors of the study's design: two districts with two cohorts within each district, and the inclusion of all of the on-file data within each cohort. By including two districts with two cohorts separated by 12 years, cohort and district effects influencing the results are moderated. However, with only two of each, cohort and district effects must be considered as viable explanations for all of the results of this study until confirmed elsewhere. Additionally, by including the entire cohorts, rather than a random sample of each cohort, internal sample bias due to random sampling is reduced, while increasing the overall number of student cases, and thus the power of the overall study. Although these methods do help with the external validity of this study, the generalizability of the findings presented here are questionable when considered in other contexts. The study demands replication.

A suggested study to follow this pilot work is to perform a similar study for multiple districts, including all students from multiple districts in a mid-sized metropolitan area. A proposed study such as this could include 10 to 20 districts, and thousands of students, with multiple cohorts from each district represented. If student grade patterning across districts and cohorts in a study such as the one proposed appears similar to the findings presented here, then this would provide a large boost to the generalizability of these findings. In addition, with more students, districts and cohorts, additional clusters of student grade patterns may be discovered. One could imagine students who do well in elementary school but experience problems beginning in middle school or high school, in addition to the four sets (high-high, high-low, low-high and low-low) detailed here.

Additionally, recent evidence has emerged in the broader cluster analysis literature for biological and bioinformatics sciences which has relevance to cluster analysis of educational data. New studies argue for the inclusion of thousands of cases, rather than the average that is used in the biological sciences which is 100 cases or less (Dolled-Filhart *et al.*, 2006; Ein-Dor *et al.*, 2006; Sima & Dougherty, 2006; Sorlie *et al.*, 2006). The study detailed presently, with 361 cases of student grades, is between the low end of cases, about 100, and the argued for 1000 or more cases in the cluster literature. This debate will surely continue in the realm of the bioinformatics literature. These methods should be replicated in a larger and broader context of districts and schools, such as the proposed study above, to determine if the results replicate in a different context. In the next sections, the research questions are revisited, in turn.

The Correlation of Grades and Standardized Assessments

This study has addressed three main research questions, and come to many conclusions. The first research question deals with the possibility that grades and standardized tests, while considered separate assessment regimes in the past, may be converging over time due to the increasing pressures from the accountability movement at both the state and federal levels, penetrating into the classroom and modifying teacher's curriculum decisions and to align with the state standardized tests. This has been discussed and hypothesized in the literature (Busick, 2000; Carr, 2000; Carr & Farr, 2000; Porter & Smithson, 2001; Shepard *et al.*, 2005; Streifer, 2004; Waters, 2000). If true, school leaders have another tool in data driven decision making through the use of grading systems that are correlated with and help predict student state assessment scores. Yet this idea that grades and standardized assessment are converging over time, and thus the correlation between the two systems is rising, has not been empirically tested using subject specific grades prior to the present study. This study presents initial evidence that appears to be mixed on this issue. For one of the districts, West Oak, it does not appear that grades and one standardized test, the ACT, are becoming more correlated, while for the other district, South Pine, the correlations, while not statistically significantly different, do appear to be increasing. This result is not surprising given the known variable nature of curriculum, instruction and assessment in schools and districts. Additionally, the entire difference or non-difference between the 1994 and 2006 cohorts in either district can be entirely explained as cohort effects, in which the 2006 South Pine students were a random occurrence of students whose grades and ACT scores correlated higher than the 1994 South Pine cohort.

Additionally, these correlation results are questionable because the state standardized assessment scores could not be used to compare the 1994 and 2006 cohorts since the test scores from the two time points were not on similar scales and the West Oak 1994 cohort only included students who graduated on time. So a less desirable standardized test, the ACT, which was given to only a subpopulation of the student sample, was used; this may have also led to a biased and erroneous result. As with all of the conclusions of this study, but especially for the hypothesis of the increasing correlation of grades and standardized assessments, the results must be replicated in a larger setting to better understand if grades and standardized assessments are converging. In addition, as discussed in chapter III, the data presented here is only baseline data for two time points. The question remains as to if the increasing correlation is a trend over time between these two time points, and if the increasing correlation is due to accountability pressures, or is due to some other influence in the school. This may only be ascertained through additional qualitative studies in which teachers and administrators are interviewed and/or surveyed to gain an understanding of what may have lead to a potential convergence of grades and standardized assessments. However, the results presented here also suggest an interesting follow up study. Since grades and ACT scores may be converging somewhat for one of the districts but not the other, these two districts may present a natural comparison for such a qualitative study. The different processes of each district and the approaches to the state standards, the ACT, and grades may be different and of interest between the two districts. A possible qualitative study of this difference could shed light on how school districts are reacting and adapting to the

accountability policies and the pressures of state curriculum frameworks and assessments.

A Success at School Factor (SSF)

The second major finding stemming from the first research question on the correlation of grades and standardized assessments is that the data presented here suggest not only that grades are useful assessments for consideration for data driven decision making by educational leaders, but also that grades might also measure a Success at School Factor (SSF). The fact that standardized assessments such as the ACT and the state's high school assessment for these two districts moderately and significantly correlate with subject specific grades is not a new finding. Past research has shown that grades and standardized assessments not only correlate (Brennan et al., 2001; Woodruff & Ziomek, 2004), but that ideally, grades should provide criterion validity for standardized tests and thus should at the least, moderately correlate with standardized tests (Linn, 1982). But though standardized tests are reported to school and district leaders, policy makers, and the press, with the implication that standardized test scores should be used to drive improvement in schools (Linn, 2000) (as codified in the NCLB legislation), grades are rarely used for data driven decision making in schools. Grades are seen as subjective and inconsistent, as demonstrated in the discussion of hodge-podge grading practices (Brookhart, 1991; Cizek, 2000; Cizek et al., 1995-1996; Cross & Frary, 1999; Linn, 1982; Shepard et al., 2005). This leads one to conclude that while the life of students and teachers revolve around compliance with and creation and assessment of grades and grading practices (Bailey, 1976; Hargis, 1990; Kirschenbaum et al., 1971; S. Simon, 1976), much of this work seems to be ignored by administrators, policy makers,

and the government in the current accountability movement. Darling-Hammond and associates, in a chapter from their recent book on what teachers should know and learn for effective instructional practice, state "there are three important audiences for grades: parents, external users such as employers and college admissions officers, and students themselves" (Shepard et al., 2005, p.298). Grades are not seen as useful data; teachers, school leaders and district personnel are absent from the "important audience" for the grades teachers assign. In stark contrast to this omission, one of the central arguments of this study is that not only are grades useful by the school in which they are created, but that grades are useful because they may be an assessment of both academic knowledge and how well a student is able to engage in the social processes of being schooled.

While the data presented here should be considered tentative, calling for replication, the results of this study suggest that when teachers assign grades, those grades are an assessment of two variables: a student's academic knowledge, and a student's ability to negotiate the social processes of school, namely a success at school factor (SSF). This was evidenced through the moderate correlation of ACT and state standardized test scores with core subject grades, but not with non-core subject grades, even as core and non-core subject grades moderately correlated with each other. These results suggest that these two sets of correlations, ACT with grades and core grades with non-core grades, explain two different variance structures in the data. Assuming that the ACT and the state standardized test assess academic knowledge, then it can be hypothesized that the moderate correlation of the ACT with core-subject grades is a measure of the academic content of those subjects, explaining about 25% of the variance in core-subject grades (correlation of about 0.5). However as the ACT did not correlate

with non-core subject grades, it might be an indication that those subjects did not contain the academic knowledge that the ACT assesses. This is corroborated with the state standardized test scores also not correlating with non-core subject grades. Conversely, core and non-core subject grades moderately correlated, indicting a similarity in the variance structures between core and non-core grades that does not exist between grades and the two standardized assessments studied. Again, cautioning that these results are preliminary it is the hypothesis here that the correlation between core and non-core grades represents a success at school factor. The findings, however may only be specific for the students studied and thus have little external validity. Also, the correlation evidence should be considered relatively weak due to the small sample sizes and the overlapping confidence intervals.

Additionally, and maybe more importantly, a SSF is also indicated by the K-12 cluster analysis data presented in chapter VI. Close inspection of the grade cluster patterns during early elementary, but even more consistently at the middle and high school levels, shows that student achievement is generally subject independent. Students who do well in one subject generally do well in all subjects across grade-levels and years of schooling, while students who score poorly in one subject generally score poorly in all subjects. This corroborates the correlation data on a SSF, a student's ability to negotiate the social processes of being schooled such as attending class, participating, and being well behaved. This is an important variable that teacher assigned subject-specific grades may assess. If grades are considered a measure of *both* academic achievement and SSF, rather than as a poor assessment of academic achievement alone, then a student's early

drop in grades may signal an important intervention point for district and school level data driven decision making.

If one combines the findings of past research on grades with the data presented here, the idea of a success at school factor (SSF) is reasonable. If one accepts the findings from the hodge-podge grading literature, that teachers use grades to assess much more than academic knowledge, including attendance, participation, homework completion, behavior, and extra credit assignments, then these factors that have been cast in a pejorative light in the past research literature on assessments may instead be useful as assessments of all of these factors combined, which would indicate a student's ability to conform to teacher, classroom and school social demands for the act of being schooled. This "hidden curriculum" (Bracey, 1994; Wood, 1994) is the hypothesized success at school factor (SSF). Additionally, a SSF may also indicate challenges faced by a student outside of school that influence that student's behavior and participation within the school building. These challenges may be family and economically based, such as if the student's family begins to undergo a period of high stress, due to the loss or switching of jobs by a student's parents, parental divorce, or other family strife. These challenges may also be student centered, arising from a behavioral or learning disability that had gone undetected.

The idea behind a SSF is not new. These issues have been well documented, showing that from an early age and then throughout the schooling process, a child's success at school depends on the child functioning well in multiple domains, including behavioral, attention, social and academic (Alexander et al., 2001; Flanagan *et al.*, 2003; Hamre & Pianta, 2001, 2005). All of these factors could contribute to an increase or

decrease in a student's willingness to participate in the social processes of school. Hence, the argument here is not for the existence of a SSF. While not articulated before as a "Success at School Factor", the point that multiple social processes must be negotiated to succeed at school is well studied and known. The point that grades may be an assessment of both academic knowledge and as a student's ability to negotiate the social processes of school has also been well detailed in the past (Parsons, 1959). However this point seems to have been lost in the grading literature, as the focus over the past forty years seems to have centered on the point that grades do not appear to be very reliable when it comes to assessing academic knowledge. This study does not attempt to address why the literature has focused so intently on the academic component of grades and ignored the social component in the recent literature. What is new on this subject for this study is the argument that grades may be an assessment of both academic knowledge and SSF. While not a new idea, it is argued here that a SSF should be re-introduced in the discussion of grades and the use of grades by researchers and practitioners. This study presents a viable way to do so.

The fact that grades appear to assess SSF is important due to the additional findings of this study that show that grade patterns are predictive of on-time graduation. Historically, standardized tests have lacked criterion validity measures that have linked high standardized test scores with on-time graduation, while grades, and specifically extremely low and failing grades, have been shown to correspond with higher rates of dropping out (Alexander et al., 2001; Allensworth, 2005; Allensworth & Easton, 2005; Montes & Lehmann, 2004; Wood, 1994). This study has confirmed and extended the findings that grades are useful predictors of student graduation. But a tension that exists

in the literature is the question as to why grades are predictive of dropping out if grades are merely subjective hodge-podge measures that do not appear to be consistent from teacher to teacher. The contention here is that grades are predictive of on-time graduation, as well as future grading patterns, because grades are an assessment of both academic knowledge and SSF. If a student has high academic aptitude but is unable to negotiate these social processes of school such as showing up to class on time, participating, doing their homework, and generally "playing the game" that is the American schooling process, that student will receive a low grade. Those low grades are predictive of future low grade trends as well as not graduating on time. This problem is compounded for a student that also lacks the academic aptitude or foundational skills in reading or mathematics that would also correspond to low scores on academic achievement tests.

The Hargris Hypothesis

In reference to the second research question proposed, this study provides evidence that supports the Hargris hypothesis, that early grade patterns predict future grade patterns (Hargis, 1990; Kirschenbaum et al., 1971). Results of cluster analysis show that early grades in elementary school are generally predictive of later grading patterns, and, by middle and high school, grade patterns are highly predictive of future grade patterns. Implicit in the Hargris hypothesis is that the assignment of early grades is the *cause* of later grading patterns. The data presented in this study is ambiguous on this issue. Hargris argues from the perspective of the teacher expectancy literature (Elashoff & Snow, 1971; Hargis, 1990; Raudenbush, 1984; Rosenthal & Jacobsen, 1969; Spitz, 1999), intimating that teacher perceptions of potential student ability in the early

elementary grades is one of the main causes of future student success or failure at school, and that those perceptions may be based on a multitude of factors outside of a student's actual ability, such as family socioeconomic status. Studies from the expectancy literature, while much critiqued as discussed above, concluded that early teacher perception of student ability, within the first few weeks of first grade, could be a major determinate of future student outcomes.

Although these data support Hargris' hypothesis, they do not provide evidence for or against the expectancy literature's hypothesized cause of these grade patterns. It may be true that teacher expectancy is the main driver of student grade patterns and that students who receive high grades in early elementary school are motivated into cycles of higher grading patterns, while students who receive low grades due to the self fulfilling prophecies of early teacher deficit thinking become locked into a cycle of low grading patterns, which is difficult to escape from. The data presented here do not provide enough evidence to judge the cause of student grade patterning.

Nevertheless, while these long-term consistent student grade patterns may be due to teacher expectancy, there is an alternative explanation: teachers instead may be very adept at assessing a student's ability to negotiate the social processes of school, a success at school factor (SSF), from an early age, and grades may be an indication of that assessment. Since we know that teacher perceptions of grades confirm the "hodge-podge" grading practices, in that grades are an assessment of the multiple social norms of schooling, such as attendance and participation, then it is reasonable to believe that rather than dooming children from an early age to patterns of low or high grades, as the expectancy literature implies, teachers are accurately assessing a student's ability to

negotiate the social processes of school, and that this assessment is reflected in the grades that a student receives. A history of high grade patterns may indicate that a student is not only acquiring academic knowledge, but also conforming to the social norms of the schooling process. Conversely a history of low grade patterns may indicate that a student is not acquiring academic knowledge and may also not be learning how to negotiate the social norms of schooling. It can be imagined that if a student has not learned how negotiate the system of school, and is not turning in homework, participating, or attending class, then that student could fall quickly behind in their academic work, which would predispose that student to lower grades and compounding problems.

Additionally, Hargris does not address the issue of students who do not conform to the overall grading pattern trends; students who may start with high grades but then their grades decrease over time, or students who start with low grades which then increase over time. As shown in data presented here, generally student grade patterns are either high to high or low to low. For some smaller clusters however, students may start elementary school with low grades across subjects, and then show improvement over time. Other clusters of students start with high grades across subjects, but then continue to loose ground over time; the high-low and low-high clusters. In contrast to the Hargris hypothesis, the hypothesis presented here of a Success at School Factor (SSF) is able to explain both sets of clusters. For students in the high-high clusters, those students may have an aptitude for both academic knowledge and SSF, and their grades reflect this. For low-low students the opposite may be true. In addition, for some students, the ability to perform within the social process of school may be a necessary skill to acquire before the acquisition of academic knowledge may take place. A logical conclusion is that for many

students the acquisition of the skills to perform well within the social norms and requirements of the schooling process is a necessary step before they are able to efficiently acquire academic knowledge.

As opposed to the Hargris hypothesis, a SSF also may explain the low-high and high-low clusters. Early elementary teacher expectancy does not adequately explain how these two clusters of students may exist. If teachers perceive early-on that a student will get high or low grades, and this expectancy is a self fulfilling prophecy, then students starting with high grades and then falling, or students starting with low grades and then rising are not well explained. If instead, grades are an assessment of a student's ability at being schooled, a SSF, then students in a low-high cluster may initially be behind the rest of their cohort in learning both the academic knowledge and the social norms of schooling. However, after a time, students may learn what is expected of them and begin to conform to the social processes of school. Additionally, these students may also be developmentally behind their cohort, and, with time, may gain the ability to learn the academic and social norm knowledge required to perform at school. It is interesting to note that the low-high cluster diverges at the second grade from the low-low cluster, and that few students in this dataset appear to "recover" in this way after elementary school. It may be that there is a short window of time in which a student may "catch-up" to peers in the cohort. If this does not happen in early elementary school, then the numbers of challenges continue to rise for the student, and they remain in the low-low cluster. This idea is supported in the dropout literature and will be further discussed below.

Conversely, the inverse may be true for the high-low cluster of students, in which they are progressing well in early elementary school, but as they reach fourth grade their

grades begin to fall. This could be due to an increase in the requirements of academic knowledge, transitioning from memorization to comprehension in both reading and mathematics, such as the change from "learning to read" to "reading to learn" (NCES, 2001). Another explanation is the additional requirements for participation and behavior as the academic press of the higher grade-levels increases and as students begin to enter puberty. Interestingly, the early learning literature concentrates much attention on both the second and fourth grades, referring to both as assessment points in which students may have individual issues that can be helped with individualized educational plans and more specific attention to their needs (Kamii & Joseph, 2003; Torgesen, 2002).

Thus, a Success at School Factor may be a better explanation for grade patterns than the Hargris hypothesis. Rather than early teacher expectations of student ability influencing a student's entire future grading pattern, a student's ability at negotiating the social processes of school could be a contributor to a student's future grade patterns as teachers accurately assess a student's SSF through grades. Rather than casting grades in a pejorative light, as much of the grading literature has done to date, instead grades may be useful as an assessment of a student's SSF. That assessment is important when considering data for decision making. A point discussed in more detail below. Additionally, the evidence presented here indicates that a student might have a limited time window in early elementary school to catch-up in either academic knowledge or SSF, and that by the end of elementary school, student grade patterns are generally set. Students may be too far behind to reasonably expect them to catch up to their cohort, arguing for early rather than later at-risk intervention strategies.

It must again be noted, however, that the data supporting a SSF is tenuous at best. These results rest on a small and intact data sample. These limitations must be taken into account when considering the veracity of the claims made here for the existence of a SSF.

Prediction of Not On Time Graduation (NOTG)

The final research question addressed by this study was the extent to which student grade patterns are predictive of qualitative student outcomes, such as graduating or not graduating on time. The results presented in chapter VI show that by using hierarchical cluster analysis to cluster the patterns of student grades, K-12 subject specific grades are useful in predicting a student's chances of not graduating on time (NOTG). This prediction method appears to be comparable to past prediction methods of students at-risk of dropping out and not graduating on time, such as the methods reported by Gleason and Dynarski (2002). Moreover, while much of the "at-risk" literature on student dropout prediction has focused on the high school level, and to a much lesser extent on the middle school level (Gleason & Dynarksi, 2002; Montes & Lehmann, 2004; Rumberger, 1995), cluster analysis of grades appears to be superior in the accuracy of predicting students at-risk of not graduating on time at the middle school and elementary levels (*see Table 18*). Furthermore, while Gleason and Dynarksi present a "best case" at-risk predictor using regression composites (2002), as discussed in chapter II, principals and school leaders rarely use regression statistics due to the complexity of regression calculations; the violation of multiple assumptions of regression by using nested, dependent and multicolinear district-level data; combined with little interest in estimating the mean of the general population when they really want to know what may happen with their students within the next few years (Creighton, 2001a). Accordingly, school leaders

rarely create regression composites of multiple student variables to predict a student's risk of dropping out. Instead they use individual variables to identify students as at-risk (Montes & Lehmann, 2004). The central point is that the method of cluster analysis of grades presented here may be more accurate, applicable, and "user friendly" in predicting students at risk of not graduating on time considering that 1) the method is comparable to past prediction methods using high school level data, and appears superior at the middle and elementary levels; 2) cluster analysis does not have the assumption violation issues of regression analysis of multicolinearity, dependency of cases, and nested levels of data and instead is made more robust when the data contains such underlying structure; 3) is applicable to entire cohort, school and district datasets rather than random samples; 4) uses grade data that is currently collected on students rather than requiring additional outside assessments; 5) and employs the use of grades, which have face validity for teachers and parents, are collected from the earliest grade-levels, and have the potential to indicate specific subjects and grade-levels for possible intervention. In sum, cluster analysis appears to provide an advance over current at-risk prediction methods. It could be used for data driven decision making by school leaders to help direct the limited resources of a school district in service to students who may be experiencing challenges in school and deserve intervention.

The overriding theme of this study is that grades are useful and predictive as assessments of student progress. It is not a novel idea that a student's ability to negotiate the social processes of school matters for that student's eventual life outcomes, such as on-time graduation.

It has been well documented in multiple studies that a student's risk of dropping out of school is not attributable to a single event, but rather appears to be a long-term process through which the accumulation of multiple challenges over time in a student's schooling career continually build-up, culminating in a student's decision to drop out (Bryk & Thum, 1989; Christenson & Thurlow, 2004; Delgado-Gaitan, 1988; Ensminger & Slusarcick, 1992; Gleason & Dynarksi, 2002; Gutman *et al.*, 2003; Jimerson et al., 2000; Randolph & Orthner, 2006; Rumberger, 1995). This process has been termed a "dynamic" or "life-course" process (Alexander et al., 2001; Jimerson et al., 2000). Furthermore, in reference to school dropouts, it has been shown that social capital, social support, and emotional support of students at all levels of schooling is important for helping students gain the skills necessary to succeed in school, and that rather than being focused on academic knowledge, much of this need for social support is centered on helping students connect with the social processes of school in an effort to minimize a student's risk of not graduating on time, and helping them to "play the game" and follow the rules of schooling (Barker, 2005; Croninger & Lee, 2001; Delgado-Gaitan, 1988; Hamre & Pianta, 2005; Knesting & Waldron, 2006; Miller, 2005; Zvoch, 2006).

These social factors all relate to the Success at School Factor hypothesized in this study, and help to support the idea that the successful negotiation of the social processes of being schooled is an important component in student's lives, since it may lead to greater participation in school, an increase in general academic achievement, and a higher probability of graduating on time. Moreover, a contention of this study is that an assessment of SSF appears to be a component of grades, and that grade patterns when

examined through cluster analysis, are useful in helping to determine students at-risk of not graduating on time, and is an advance over current methods of at-risk prediction.

While this study presents a novel method and use of teacher assigned, subject specific grades in predicting students at-risk of not graduating on time, it does not address the issue of what should be done once students are identified. While outside the scope of this study, it is important to address this question since accurate identification is only the first step of many in helping to address the needs of students who may be experiencing difficulties with school. However, to date, little work has been done to systematically evaluate at-risk prevention programs.

For most of the evidence, methodological problems persist which inhibit a robust evaluation of what works, such as biased groupings and estimates of effects, since randomized controlled trials are rarely performed in this area (Agodini & Dynarksi, 2004; Lehr et al., 2003). Nevertheless, what the literature indicates is that historically, most dropout prevention programs appear to not reduce student dropouts (Dynarski & Gleason, 2002). As reviewed by Dynarski and Gleason (2002) and Lehr *et al* (2003), these programs mostly occur at the high school level and consist of helping students build self-esteem, overcome personal and family issues and increase attendance through periodic counseling; consist of the creation of smaller school settings; or provide tutoring or mentoring services. Similar programs at the middle school level have had somewhat more of an impact, but as discussed above, the accuracy of identification of students at risk of dropping out using middle school level data has been low and problematic to date. Hence, any program that appears to work using middle school level data, may have "worked" only to the extent that the majority of the students identified for at-risk

interventions were mis-identified originally as being students at risk of dropping out.

Acknowledging that much more high-quality work is needed in the evaluation of dropout prevention programs before any one individual program can be recommended over another (Dynarski & Gleason, 2002; Lehr et al., 2003), recent literature has begun to urge for a shift from a deficit model of attempting to prevent dropouts, to a more positive model of promoting and encouraging successful school completion (Christenson & Thurlow, 2004). From the perspective of the results presented here in chapter VI, dropout prevention programs should focus more on the earlier grade levels, rather than almost exclusively at the high school level. To this end, a recent study showed that first-grade students with known characteristics of school dropout taught in classrooms with multiple dimensions of support (including behavioral, attention, academic and social) increased scored higher on academic and social achievement scales, than comparable children who attended classrooms with less supportive environments (Hamre & Pianta, 2005). Interventions such as this, which provide early assistance for *both* the academic and social needs of children, provide an attractive future avenue for intervention studies and for district strategies to help students learn and ultimately graduate on time.

Cluster Analysis of Subject-Specific Teacher Assigned Grades

Chapter VI presents a novel application of cluster analysis for the study and use of subject-specific teacher assigned grades. Patterns of longitudinal student grades appear to be predictive of future student grades and qualitative outcomes, such as on-time graduation. In addition, specific sub-clusters of student grade patterns suggest early intervention points in student's careers in schools for students who may be at risk of low school performance and eventually not graduating on-time.

Cluster analysis has been rarely used in education to date. This may be due to the perception that cluster analysis, a descriptive multivariate statistic, is a less sophisticated procedure for statisticians than other multivariate statistics such as linear regression, due to the point that significance tests and confidence intervals are not readily applicable to cluster analysis (Lorr, 1983; Romesburg, 1984). However, if one conceives of school and district-level data as large and historically untapped databases that are highly multicolinear, interdependent, and nested, then cluster analysis as a data mining procedure becomes more attractive for researchers and practitioners faced with the avalanche of student-level data now collected on students at every level. The attractiveness of cluster analysis rises further when considering that these same datasets are problematic for use in regression statistics, due to these same issues with multicolinearity and dependence of cases. Student data is messy and complex. Cluster analysis can bring order and structure to that data, revealing previously unknown patterns in an effort to help drive decision making based on that data.

Combining cluster analysis with an Eisenplot in the analysis of educational data, as detailed in the methods and chapter VI, is also a novel application of this method. The majority of quantitative methods rely on aggregation of data to the mean, and the reporting of a generalized trend. For large scale studies that wish to estimate the mean of a population of students in a state or a nation, generalization to the mean is desired. However, at the school and district level, reducing data trends to the mean necessarily requires the loss of information and an increase in the theoretical "distance" between the generalized trends and the individuals for whom the data could be used for decision making (Hayman *et al.*, 1979). To tease out overall patterns and trends, this loss of data

in exchange for an overall mean has been deemed as acceptable in the literature, and newer statistical procedures have been created to recover or control for specific trends in data through deviation from the mean in many high-level statistical procedures, including all forms of regression analysis. However, these procedures all come with additional issues of controlling for assumption violations in a dataset. Nevertheless, the fine granularity of a dataset is lost with traditional inferential statistics as the statistics aggregate to the mean. This becomes extremely important when considering the use of this data for decision making for individual schools and districts. For a large enough dataset, each individual's data in any situation should theoretically be unique. Hence, if decisions are to be made about individuals based on their data, especially high stakes decisions in settings such as education, decisions should be based on the entire pattern of data of an individual to date in the system leaving the individual data points intact and available for review and alternative analysis.

At the other end of the spectrum, far removed from aggregating all data to the mean, is the practice of relying on individual data points to make high-stakes decisions, often witnessed in education as students are assigned to at-risk pull-out programs, retention, or remedial services based on one, or just a few data points, such as a single grade, test, or categorical variable (Coburn & Talbert, 2006; Creighton, 2001a). The logical middle-ground between these two extremes of generalizing to the mean or basing decisions on individual datapoints, is to acknowledge the qualitative literature and strive to produce deep and rich datasets that begin to bring together the best of quantitative and qualitative theories of knowing, bridging the divide and blurring the lines between the two. Finding a middle ground between quantitative generalizable statistical findings and

qualitative context-localized deep descriptions has been much discussed in the literature (Madey, 1982; Onwuegbuzie & Leech, 2005; Shaffer & Serlin, 2004). Cluster analysis with the inclusion of an Eisenplot is a start down this path from the quantitative side.

While the cluster analysis procedure detailed here does use means and correlations, it begins to bridge these divides between the loss of data to the mean versus examination of single data points, as well as the divide between quantitative generalized data and qualitative context-localized data, in four main ways. First, cluster analysis employs the use of numerical datasets, and is thus considered a quantitative method (Lorr, 1983; Rencher, 2002; Romesburg, 1984). Second, cluster analysis preserves the entire list of all cases, rather than aggregating all cases into a single mean, reordering a list and giving it a taxonomic structure that places each case proximal to other similar cases in the list based each case's data pattern. For schools, rather than aggregating achievement data to a mean for the school or district, cluster analysis preserves the list of cases that would go into such a mean, and gives the order of the list meaning. Third, a dendrogram, or cluster tree, allows one to visualize the organization of clusters and magnitude of similarity between clusters, revealing more about each case rather than less. Fourth, with the inclusion of an Eisenplot, every datapoint for each case for each variable is displayed in a context based on the similarity of each case's data pattern to each other case's data pattern in the dataset. This provides for a deep and rich display which makes obvious and disaggregates every datapoint used in the cluster analysis, revealing and maintaining the data of each individual while allowing for pattern recognition. In this way, cluster analysis is a quantitative method that employs some of the aspects of a

qualitative method in creating a deeper and thicker description. It is a visual and intuitive new data analysis tool for educators.

Cluster analysis requires the standardization, or z-scoring, of all variables within a dataset, to compensate for the overweighting in the dataspace that one variable may have that would distort the clustering patterns (Lorr, 1983; Romesburg, 1984). Z-scoring allows then for a more "apples to apples" comparison. Moreover, in clustering grade data for this study, the necessity to z-score all of the data provided unexpected benefits. First, because each subject-specific grade variable is normalized through z-scoring, grade inflation is controlled for. Second, grade data has rarely been z-scored in the literature, however z-scoring of grades, especially at the early elementary stage, may be an important innovation. If one considers that it has been shown that low or failing grades at the high school level are predictive of student dropout (Alexander et al., 2001; Allensworth, 2005; Allensworth & Easton, 2005), but that the distribution of grades may be narrower and skewed towards higher grades in early elementary school (no students failed any subject at any grade-level before 6th grade in the dataset presented here) then examining grades as a z-scored distribution rather than as a fixed scale is important. As shown in chapter VI, low grades as early as first and second grade are generally predictive of future student grade patterns. However, since the grades are z-scored, "low" is relative to the distribution of each subject-specific and grade-level variable. This results in the ability of cluster analysis to reveal "low grading" patterns that would not be readily apparent if the grades were clustered based on the 4-point grading scale, such as those in the lower cluster of Figure 19, in which the low grades at the early elementary level may only be as low as a B or a C. As shown in Figure 25 in the examination of the

four clusters of high-high, high-low, low-high and low-low, a small change in overall grades at the elementary level, from a B+ to B, is significant in predicting the future outcomes of the students in the cluster, and most likely has gone unexamined in the past. Z-scoring of grades has resulted in the conclusion that not only does it appear that students with low grades in early elementary school appear to continue to receive low grades throughout the rest of their career in a district, but that "low grades" should be defined as students who skew more towards a -1 standard deviation across a standardized graded-subject variable, rather than on absolute grades, such as a C, D or F.

Cluster analysis in combination with an Eisenplot provides an additional innovation when examining longitudinal subject-specific grades of providing a visual method to assess the course enrollment patterns of all students across a dataset. In the past, a method has not been readily available which would allow for school leaders to disaggregate and examine all of the course taking patterns of all of their students, and compare the differences in those patterns between students who appear to succeed in school and those who do not. The method presented here allows one to do just that. As mentioned in the presentation of the results for Figure 19 in chapter VI, by examining the patterns of columns of data at the high school level for students in the upper cluster in comparison to the lower cluster, students in the lower cluster appear to take fewer classes overall than students in the higher cluster, and they are awarded lower grades for those fewer courses. Additionally, because each grade-level in the cluster analysis contains a repeating order of subjects from left to right (from core subjects such as mathematics, English and science, to more non-core subjects, such as band, physical education and art) "columns" of contiguous data patterns running vertically can be identified for students in

the upper cluster throughout high school, beginning to dissolve into more non-core subjects only by grade 12. Conversely, for students in the lower cluster, not only is it evident by the color-block pattern that they are taking fewer courses than the upper cluster, but because there is more "scatter" in their patterns across the repeating pattern of subjects, the students in the lower cluster enrolled in many more non-core courses than students in the upper cluster. This is a significant finding considering the research to date on the higher rate of success, graduation and college attendance for students who take core courses throughout their careers in high school (Adelman, 1999; Ayalon, 2006; Gamoran & Hannigan, 2000; Girotto & Peterson, 1999; Meyer, 1999; Trusty, 2002; Woods, 1995).

Grades and Data Driven Decision Making - Conclusion

This study has shown evidence that grades are useful when considering data driven decision making, that grades and standardized assessments may be converging over time for one of the two districts, and that cluster analysis is a new and useful method for analyzing patterns of student data to predict future outcomes. This analysis may be an advance over past practices that is more useful, has fewer assumption violations, and has more face validity than past methods. The literature to date on data driven decision making indicates that when teachers and school leaders collaborate around student-level data with a focus on improvement of educational practice, the process of open communication, dialogue and a focus on student's performance to date in the system is helpful in encouraging school success and an increase in professional collaboration amongst the staff (Bernhardt, 2004; Coburn & Talbert, 2006; Halverson *et al.*, 2005; Kerr *et al.*, 2006; Thorn, 2002; Wayman & Stringfield, 2006a, 2006b; V. M. Young, 2006).

The contention of this study is that grades, and analysis of grades through cluster analysis, may be useful for data driven decision making in schools and school districts. Rather than have schools make two copies of a report card and send one home to the parents and one into storage in the basement, school leaders should bring that data back up out of the basement and put it to use for data driven decision making for multiple reasons. First, grades are already generated as part of the system, and so in a way they could be seen as "free" or low cost, especially in comparison to the current movement in many districts across the nation discussed in chapter II to add increasing levels of periodic assessments to help predict state assessments, spending both money and instructional time on what may be unnecessary additional test preparation. Second, grades appear to be predictive of future student grade patterns and on-time graduation, and for one district in this study, the correlation between grades and a standardized assessment may be rising over time. Hence, rather than ignore the grading system, which schools already devote enormous amounts of time to generating, that data can be used more efficiently by including it in the data driven decision making process, to analyze the performance of each student, predict future performance, and help direct the limited resources of a school district to students who could most benefit. Third, the method presented here utilizes data that schools already possess, and mirrors what could be considered a "typical" district dataset. Fourth, because this method uses data that is already present in every school district, the two largest hurdles to practitioners using grades and cluster analysis are the extensive amount of effort required to input grades into an electronic database and teaching practitioners how to conduct and read cluster analysis and Eisenplot outputs. With the continuing increase in district use of electronic

databases to store all of their data (Streifer, 2002; Wayman, 2005; Wayman & Stringfield, 2006b; Wayman et al., 2004), the issue of transferring grade data from paper report cards into an electronic database disappears. Additionally, since cluster analysis requires much less attention to the traditionally problematic issues of multicolinarity and case dependence of regression analysis, analysis by district leaders using clustering and Eisenplots should be less difficult than most other statistical analyses once they are trained on how to read a cluster tree and an Eisenplot. And fifth, since grades have face validity with teachers, parents and students, using grades in addition with standardized tests for data driven decision making may help increase buy-in on data-based decisions from these multiple stakeholders.

The final point is that analysis of long-term grading patterns should be considered the job of school leaders and district central office staff, not teachers, thus making this analysis an administrative and leadership issue. A teacher's day is already full, and adding the requirement to cluster or pattern their students by classroom would add unnecessary work. Additionally, a teacher must be concerned with her entire class and the near-term needs of all of her students, working to improve daily instruction for tomorrow. It is the job of school and district administrators to provide the data analysis for teachers so that they may see the connections in their practice throughout a school system and how each teacher's practice influences the outcomes for students over time. To keep from burdening teachers with an ever increasing array of responsibilities, school leaders must provide the finished analysis for discussion, rather than requiring teachers to perform the analysis themselves. Also, since the addition of data only increases the granularity of clusters within a clustered dataset, it would be unreasonable to ask

individual teachers, or even individual schools to cluster their data. Rather, districts should cluster all of their data to create the largest and most robust dataset available. The methods detailed here provide a means to focus in on the long-term district-wide trends of student grade patterns, and at the same time pinpoint specific time and data points for interventions. If a timepoint, subject, grade level, or specific cluster of students is identified as in need of assistance, that assistance would take political power and financial backing to implement given the limited resources of a school district. Only central office staff and school leaders have such power, as well as a school and district-wide vision that could be enhanced through the addition of grades and cluster analysis to ongoing efforts at data driven decision making.

APPENDICES

Bowers, A.J. (2007)

APPENDIX A

Table 19: *Course names and percentages of students who attended each specific course for each subject grouping during 10th grade semester 2, full dataset***.**

Bowers, A.J. (2007)

Mathematics Class Name	% of enrolled students	English Class Name	% of enrolled students	Science Class Name	% of enrolled students
Not Enrolled	24.7%	English 10 Taken	28.3%	Biology	36.8%
Int Math II	12.2%	Not Enrolled	23.9%	Not Enrolled	25.2%
Geometry	7.8%	Lang Arts II	19.7%	Earth Sci	11.9%
Math II	7.8%	English II R	5.3%	Env Sci	8.0%
Math I	6.9%	Hon Eng 10	4.4%	Earth Science	5.0%
Algebra B	5.3%	English 10 Honors	4.2%	Practical Biology	4.2%
Pre Alg	5.3%	English II H	3.1%	Life Sci	1.1%
Algebra I	3.3%	English 9	1.7%	Anat/Phys	0.8%
Algebra II	3.1%	Lang Arts III	1.4%	Physical Science	0.8%
App Math II	2.8%	Eng 10	0.8%	Life Science	0.6%
Shop Math	2.5%	English II	0.6%	Basic Chemistry	0.3%
Algebra A	2.2%	ESL - English 3	0.6%	Basic Earth Science	0.3%
Int Math II-H	1.9%	IEP English	0.6%	Basic Human Science	0.3%
App Math I	1.4%	Lang Arts III H	0.6%	Biology 1	0.3%
Basic Geometry	1.1%	Basic English	0.3%	Biology 2	0.3%
Int Math III	1.1%	Basic English 2	0.3%	Chem in the communit	0.3%
App Math III	0.8%	British Lit	0.3%	Chemistry	0.3%
ESL - Math	0.8%	Composition	0.3%	ESL Science	0.3%
Math	0.8%	English	0.3%	Fd Biology	0.3%
Algebra 1	0.6%	English 10B	0.3%	Gen Bio Sci	0.3%
Applied Math	0.6%	English 2S	0.3%	Gen Sci	0.3%
IEP Math	0.6%	English Lang Studies	0.3%	General Science I	0.3%
Int Math III-H	0.6%	English Skills	0.3%	Hon Eng 9	0.3%
Pre-Algebra	0.6%	English V	0.3%	Phy/Earth Science	0.3%
Alg 2nd half	0.3%	ESL - English 2	0.3%	Phyiscal Science	0.3%
Alg Essentials II	0.3%	ESL - English A	0.3%	Phys/Earth Science	0.3%
Alg I	0.3%	IEP Eng	0.3%	Physical/Earth Sci	0.3%
Alg II/Trig	0.3%	IL Lit 10	0.3%	Physics	0.3%
Algebra	0.3%	Lang Arts I	0.3%	Sci Concept 2	0.3%
Algebra 1-2	0.3%	Language Arts 10	0.3%	Y Science 2A &B	0.3%
Basic Geom	0.3%	Look at Lit	0.3%		
Cons Math	0.3%	Y English 2A &B	0.3%		
Cons. Math	0.3%				
Consumer Math	0.3%				
E-Basic Math	0.3%				
General Math	0.3%				
IEP Math II	0.3%				
Int Math	0.3%				
Int Math I	0.3%				
Int Math-H	0.3%				
Integ Math 1	0.3%				
Math 9	0.3%				
Res Math	0.3%				

Foreign Language Class Name	% of enrolled students	Social Studies Class Name	% of enrolled students	Government Class Name	% of enrolled students
Not Enrolled	79.4%	US History	44.0%	Not Enrolled	94.7%
Spanish II	10.8%	Not Enrolled	29.0%	Gov/Cons Econ	2.5%
Spanish I	5.3%	Wld Hist	18.1%	Government	1.7%
Spanish 2/3	1.4%	State History	2.5%	Am Government	0.3%
Spanish 1	0.8%	Psychology	1.4%	American Govt	0.3%
Spanish III	0.6%	ESL US History	0.6%	Geography/Econ/Civic	0.3%
French 1	0.3%	World History	0.6%	Pratical Law/Econ	0.3%
French I	0.3%	Am Wrld Studies	0.3%		
French II	0.3%	Amer. History	0.3%		
German I	0.3%	Cont Global	0.3%		
Spanish I H	0.3%	Contempt WWld	0.3%		
Spanish I-H	0.3%	ESL History	0.3%		
		ESL U.S. History	0.3%		
		Global Issues	0.3%		
		History	0.3%		
		IEP Hist	0.3%		
		IEP US Hstory	0.3%		
		Mod World History	0.3%		
		Soc Studies	0.3%		
		Sociology	0.3%		
		Y-Geography 2A & 1B	0.3%		

Economics Class Name	% of enrolled students	Band Class Name	% of enrolled students	Physical Education Class Name	% of enrolled students
Not Enrolled	93.9%	Not Enrolled	84.4%	Not Enrolled	67.2%
Intro to Bus	1.4%	Band	5.3%	Strg/Cond	7.8%
Economics	1.1%	Band-HS	2.8%	Health	7.2%
Acct	0.8%	Concert Choir I	1.4%	P.E.	3.9%
Accounting	0.6%	Jazz Band	1.4%	Phys Ed	3.6%
Prin of Mkting	0.6%	Choir-HS	1.1%	Adv PE	3.3%
Retailing	0.6%	Symphonic Band	1.1%	Advanced PE	1.1%
Accounting I	0.3%	Concert Choir	0.8%	PE	1.1%
Accounting III & IV	0.3%	Choir	0.6%	Physical Edu	1.1%
Bits Business and fi	0.3%	Conc Choir	0.3%	Team Sports	1.1%
Intro Bus	0.3%	Concert	0.3%	PE Swim	0.6%
		GHS Singers	0.3%	ADV PE	0.3%
		Men's Choir	0.3%	Advaned PE	0.3%
				Dance Fitness	0.3%
				Life Rec Sports	0.3%
				PE 10	0.3%
				Phys Educ	0.3%
				Team Sports/Health	0.3%

Computers Class Name	% of enrolled students
Not Enrolled	75.0%
Bus Tech I	8.9%
Computer App	4.4%
Comp App	2.8%
Computer2/Careers 2	1.9%
Bus Tech II	1.7%
Comp Multimedia	1.4%
Computer1/Careers 1	1.1%
Bits ATC	0.3%
Cadet Media	0.3%
Com App	0.3%
Com Multimedia	0.3%
Comp/Multimedia	0.3%
Computer Apps	0.3%
Computer Pro	0.3%
Computer1/Careers1	0.3%
Health	0.3%
Intro Info Processin	0.3%

Life Skills Class Name	% of enrolled students
Not Enrolled	76.3%
Mach Woods	5.4%
Arch Draw	2.9%
Typing	2.6%
Mech Draw	2.3%
Bench Woods	1.1%
Gen Metals	1.1%
NTR/WP	1.1%
Gormet Food	0.9%
Personal Living	0.6%
Woodshop I	0.6%
Adv Mech Drawing	0.3%
Arch Drawing	0.3%
Basic Foods	0.3%
Basic Skills	0.3%
Bench Metal	0.3%
Cadet Media	0.3%
Cadet Teacher	0.3%
Communication Arts	0.3%
Coop Voc	0.3%
Drafting	0.3%
Driver Ed	0.3%
Human Resources Admi	0.3%
Ind Living Skills	0.3%
Keyboarding	0.3%
Keyboarding 1	0.3%
Mech Drawing	0.3%
Retailing	0.3%
Study Skills	0.3%

Art Class Name	% of enrolled students
Not Enrolled	82.4%
Art 2/3	4.2%
Art Foundations	4.2%
Art	1.7%
Art II	1.7%
Illustration I	1.1%
Acting Theater	0.8%
Art I	0.8%
Studio Art	0.6%
Acting Theatre	0.3%
Art Studio	0.3%
B. Theater	0.3%
Comp/Art Skills	0.3%
Drama	0.3%
Draw 1	0.3%
Illustration I H	0.3%
Intro to Art	0.3%
Theater	0.3%

APPENDIX B

Fisher confidence intervals were calculated for each correlation of grades to ACT in Figures 14 and 16 through 18. Confidence intervals were calculated as previously described (Howell, 2002). Fisher *r* to *z* transformations were estimated using equation 8 and back transformed using equation 9.

$$z_r = \frac{1}{2}\left[\log_e(1+r) - \log_e(1-r)\right]$$

Equation 8

$$r_z = \left[\left(e^{(2z)} - 1\right)/\left(e^{(2z)} + 1\right)\right]$$

Equation 9

Pairs of confidence intervals (lower and upper) are listed in Table 20 for each correlation in Figures 14, and 16 through 18. All of the confidence intervals overlap, indicating little to no statistical difference between the correlations.

Table 20: *Fisher confidence intervals for correlation comparisons*

	COMP		MATH		ENG		READ		SCI	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
West Oak										
HSGPA 1994	0.165	0.835	-0.025	0.800	0.464	0.915	0.145	0.833	0.140	0.830
HSGPA 2006	0.385	0.800	0.435	0.825	0.175	0.675	0.340	0.785	0.190	0.715
South Pine										
HSGPA 1994	0.295	0.770	0.275	0.760	0.007	0.620	0.145	0.700	0.315	0.780
HSGPA 2006	0.390	0.785	0.240	0.715	0.315	0.755	0.375	0.780	0.200	0.695
West Oak 1994										
Math HS GPA	0.075	0.809	-0.125	0.727	0.265	0.867	-0.087	0.744	0.058	0.802
Eng HS GPA	0.206	0.850	-0.223	0.676	0.540	0.929	0.282	0.871	-0.042	0.764
Sci HS GPA	0.206	0.850	-0.103	0.737	0.440	0.909	0.119	0.823	0.057	0.802
West Oak 2006										
Math HS GPA	0.130	0.677	0.272	0.755	0.019	0.619	-0.021	0.594	-0.051	0.574
Eng HS GPA	0.334	0.778	0.311	0.772	0.086	0.659	0.247	0.743	0.262	0.749
Sci HS GPA	0.235	0.732	0.312	0.773	-0.014	0.598	0.203	0.721	0.156	0.697
South Pine 1994										
Math HS GPA	0.111	0.679	0.246	0.747	-0.123	0.531	0.012	0.622	0.100	0.673
Eng HS GPA	0.250	0.749	0.140	0.694	0.111	0.679	0.161	0.706	0.165	0.708
Sci HS GPA	0.246	0.747	0.317	0.779	-0.074	0.566	0.087	0.666	0.419	0.822
South Pine 2006										
Math HS GPA	0.473	0.824	0.455	0.816	0.291	0.741	0.403	0.793	0.323	0.757
Eng HS GPA	0.339	0.764	0.156	0.670	0.328	0.759	0.319	0.755	0.091	0.632
Sci HS GPA	0.432	0.806	0.361	0.774	0.324	0.757	0.333	0.761	0.283	0.737
West Oak 1994										
Math 10 S2	-0.251	0.704	-0.216	0.722	-0.119	0.766	-0.281	0.687	-0.452	0.568
Eng 10 S2	0.107	0.819	-0.334	0.605	0.231	0.857	0.194	0.847	-0.106	0.736
Sci 10 S2	0.246	0.871	-0.114	0.750	0.175	0.852	0.125	0.837	0.000	0.796
West Oak 2006										
Math 10 S2	-0.063	0.566	0.193	0.722	-0.085	0.559	-0.193	0.478	-0.206	0.467
Eng 10 S2	0.181	0.710	0.027	0.630	-0.108	0.542	0.202	0.726	0.107	0.677
Sci 10 S2	-0.054	0.571	0.069	0.655	-0.255	0.426	-0.030	0.595	-0.111	0.540
South Pine 1994										
Math 10 S2	0.012	0.622	0.179	0.715	-0.076	0.564	-0.205	0.468	-0.054	0.579
Eng 10 S2	0.189	0.719	0.057	0.649	0.003	0.616	0.119	0.684	0.013	0.622
Sci 10 S2	0.135	0.698	0.261	0.760	-0.018	0.611	0.011	0.629	0.025	0.637
South Pine 2006										
Math 10 S2	0.254	0.723	0.319	0.755	0.177	0.681	0.190	0.689	0.085	0.628
Eng 10 S2	0.498	0.834	0.253	0.722	0.507	0.837	0.403	0.793	0.287	0.739
Sci 10 S2	0.484	0.828	0.350	0.769	0.460	0.818	0.389	0.787	0.310	0.750

APPENDIX C

Bowers, A.J. (2007)

Figure 27: *Eisenplot of hierarchical clustering of teacher assigned subject-specific grades, K-6.*

Cluster analysis of student grades (following page) indicates that for the K-6 dataset, student grade patterns cluster into two main clusters, those who eventually graduate on-time, and a high percentage of students who do not graduate on time (NOTG). Each student is aligned along the vertical axis, with subjects by grade-level aligned along the horizontal axis. **This figure is presented in color.** Z-scored student grades are represented by a heat map, with higher grades indicated by an increasing intensity of red, lower grades indicated by increasing intensity of blue, the mean indicated by grey, and white indicates no data (center). Hierarchical clusters are represented by a dendrogram (left), with a scale in standard deviation units for the clusters across the hyperdimensional dataspace (bottom left). The dichotomous categorical variables of NOTG is represented by black bars (right). The dashed green line through the center heat map indicates the division line between the two major clusters in the K-6 dataset (center). School and grade-level is indicated along the top horizontal axis (center top). Grade level increases left to right, starting with Kindergarten (K), then Elementary includes grades 1, 2, 3, 4, 5, and 6. Within each grade-level, subjects are listed in a repeating pattern as follows: K – mathematics, speaking, writing, reading; Elementary - 1^{st}-5^{th} – mathematics, reading, writing, spelling, handwriting, science, social studies; 6^{th} – reading, mathematics, English, science, band, social studies, physical education, art.

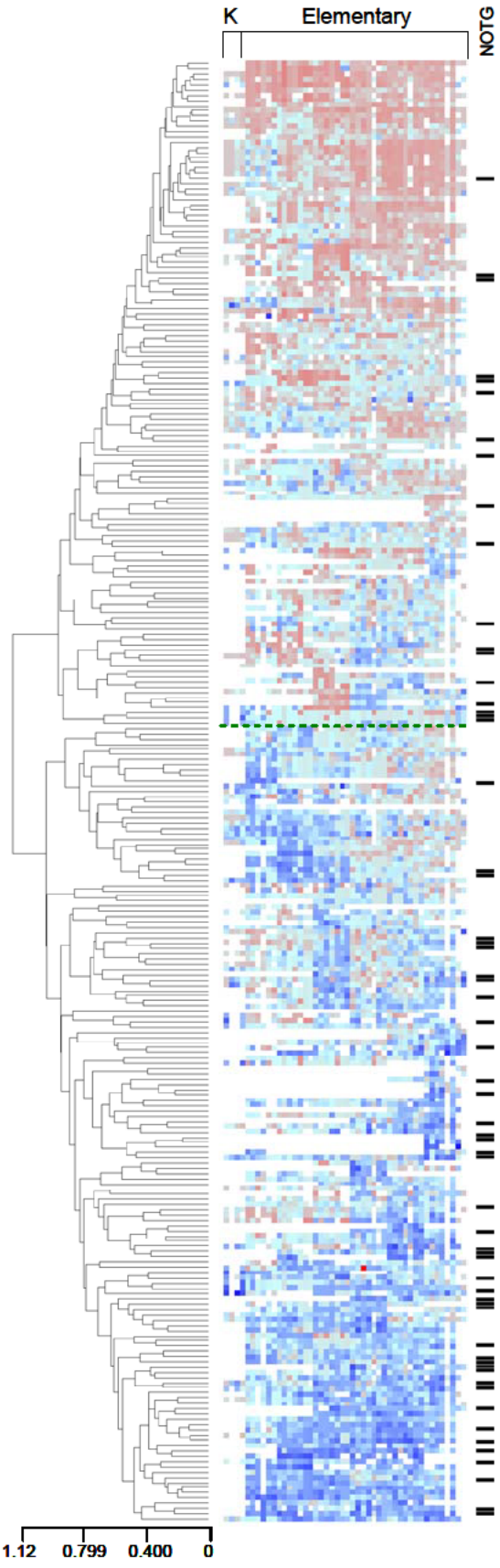
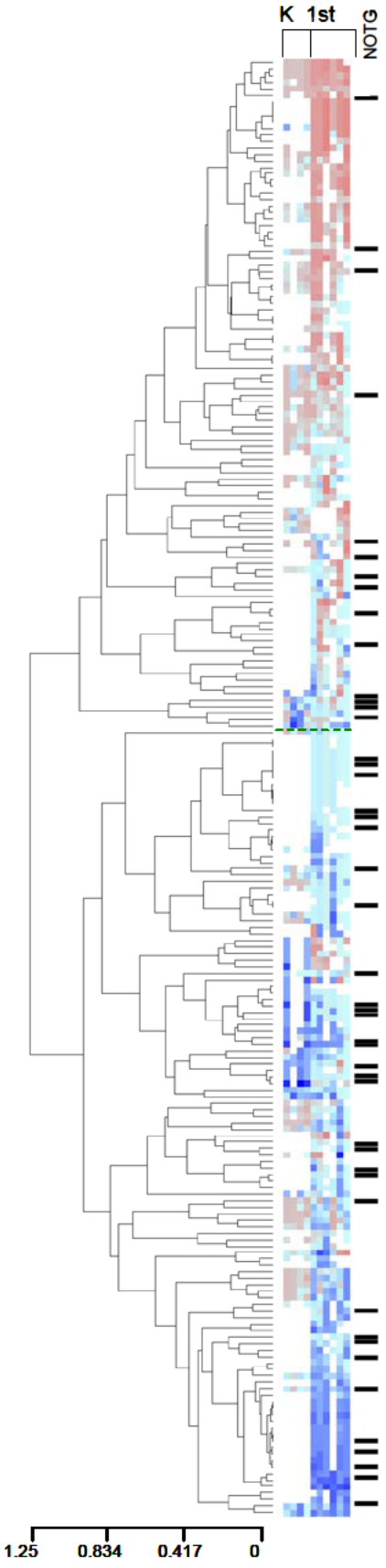

K
Elementary
NOTG
1.12
0.799
0.400
0

Figure 28: *Eisenplot of hierarchical clustering of teacher assigned subject-specific grades, K-1.*

Cluster analysis of student grades (following page) indicates that for the K-1 dataset, student grade patterns cluster into two main clusters, those who eventually graduate on-time, and a high percentage of students who do not graduate on time (NOTG). Each student is aligned along the vertical axis, with subjects by grade-level aligned along the horizontal axis. **This figure is presented in color.** Z-scored student grades are represented by a heat map, with higher grades indicated by an increasing intensity of red, lower grades indicated by increasing intensity of blue, the mean indicated by grey, and white indicates no data (center). Hierarchical clusters are represented by a dendrogram (left), with a scale in standard deviation units for the clusters across the hyperdimensional dataspace (bottom left). The dichotomous categorical variables of NOTG is represented by black bars (right). The dashed green line through the center heat map indicates the division line between the two major clusters in the K-1 dataset (center). School and grade-level is indicated along the top horizontal axis (center top). Grade level increases left to right, starting with Kindergarten (K), then Elementary includes grade 1. Within each grade-level, subjects are listed in a repeating pattern as follows: K – mathematics, speaking, writing, reading; Elementary - 1st: mathematics, reading, writing, spelling, handwriting, science, social studies.


K
1st
NOTG
1.25
0.834
0.417
0

BIBLIOGRAPHY

ACT. (2007). Act, Inc: Educational/career planning and workforce development. Retrieved January 5, 2007, from www.act.org

Adelman, C. (1999). *Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment*. Washington, DC: U.S. Department of Education.

Agodini, R., & Dynarksi, M. (2004). Are experiments the only option? A look at dropout prevention programs. *The Review of Economics and Statistics, 86*(1), 180-194.

Airasian, P. W. (1994). *Classroom assessment*. New York: McGraw-Hill Inc.

Alexander, K. L., Entwisle, D. R., & Kabbani, N. S. (2001). The dropout process in life course perspective: Early risk factors at home and school. *The Teachers College Record, 103*(5), 760-822.

Allensworth, E. M. (2005). Graduation and dropout trends in Chicago: A look at cohorts of students from 1991 through 2004. Retrieved July 7, 2006, from www.consortium-chicago.org/publications/p75.html

Allensworth, E. M., & Easton, J. Q. (2005). The on-track indicator as a predictor of high school graduation. Retrieved July 7, 2006, from www.consortium-chicago.org/publications/p78.html

Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.

Ayalon, H. (2006). Nonhierarchical curriculum differentiation and inequality in achievement: A different story or more of the same? *Teachers College Record, 108*(6), 1186-1213.

Bailey, W. J. (1976). A case study: Performance evaluation at concord senior high school. In S. B. Simon & J. A. Bellanca (Eds.), *Degrading the grading myths: A primer of alternatives to grades and marks* (pp. 74-82). Washington D.C.: Association for Supervision and Curriculum Development.

Barker, K. S. (2005). *Overcoming barriers to high school success: Perceptions of students and those in charge of ensuring their success.* Unpublished Dissertation, University of Texas San Antonio, San Antonion, Texas.

Bernhardt, V. (2004). *Data analysis for continuous school improvement*. Larchmont: Eye on Education.

Bisesi, T., Farr, R., Greene, B., & Haydel, E. (2000). Reporting to parents and the community. In E. Trumbull & B. Farr (Eds.), *Grading and reporting student*

progress in an age of standards (pp. 157-183). Norwood: Christopher-Gordon Publishers.

Bowers, A. J., Stanton, R., & Boylan, J. F. (2000). *Identification of differentially expressed mRNAs in moderately differentiated human squamous cell carcinomas.* Paper presented at the American Association of Cancer Research annual meeting, San Francisco, CA.

Bracey, G. W. (1994). Grade inflation? *Phi Delta Kappan, 76*(4), 328.

Brennan, R. T., Kim, J., Wenz-Gross, M., & Siperstein, G. N. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts comprehensive assessment system (MCAS). *Harvard Educational Review, 71*(2), 173-215.

Brookhart, S. M. (1991). Grading practices and validity. *Educational Measurement: Issues and Practice, 10*(1), 35-36.

Brunner, C., Fasca, C., Heinze, J., Honey, M., Light, D., Mandinach, E., et al. (2005). Linking data and learning: The grow network study. *Journal of Education for Students Placed at Risk, 10*(3), 241-267.

Bryk, A. S., & Thum, Y. M. (1989). The effects of high school organization on dropping out: An exploratory investigation. *American Educational Research Journal, 26*(3), 353-383.

Busick, K. (2000). Grading and standards-based assessment. In E. Trumbull & B. Farr (Eds.), *Grading and reporting student progress in an age of standards* (pp. 71-86). Norwood: Christopher-Gordon Publishers.

Cameron, S. V., & Heckman, J. J. (1993). The nonequivalence of high school equivalents. *Journal of Labor Economics, 11*(1), 1-47.

Campbell, L. (2004). As strong as the weakest link: Urban high school dropout. *The High School Journal, 87*(2).

Carr, J. (2000). Technical issues of grading methods. In E. Trumbull & B. Farr (Eds.), *Grading and reporting student progress in an age of standards* (pp. 45-70). Norwood: Christopher-Gordon Publishers.

Carr, J., & Farr, B. (2000). Taking steps toward standards-based report cards. In E. Trumbull & B. Farr (Eds.), *Grading and reporting student progress in an age of standards* (pp. 185-208). Norwood: Christopher-Gordon Publishers.

Christenson, S. L., & Thurlow, M. L. (2004). School dropouts: Prevention considerations, interventions, and challenges. *Current Directions in Psychological Science, 13*(1), 36-39.

Cizek, G. J. (2000). Pockets of resistance in the assessment revolution. *Educational Measurement: Issues and Practice, 19*(2), 16-33.

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. E. (1995-1996). Teachers' assessment practices: Preparation, isolation and the kitchen sink. *Educational Assessment, 3*(2), 159-179.

Coburn, C. E., & Talbert, J. E. (2006). Conceptions of evidence use in school districts: Mapping the terrain. *American Journal of Education, 112*(4), 469-495.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction and research. *Educational Evaluation and Policy Analysis, 25*(2), 119-142.

Coleman, J., Campbell, E., Hobsen, C., McPartland, J., Mood, A., Weinfeld, F., et al. (1966). *Equality of educational opportunity survey*. Washington, D.C.: U.S. Government Printing Office.

Creighton, T. B. (2001a). Data analysis and the principalship. *Principal Leadership, 1*(9), 52-57.

Creighton, T. B. (2001b). *Schools and data: The educator's guide for using data to improve decision making*. Thousand Oaks: Corwin Press.

Croninger, R. G., & Lee, V. E. (2001). Social capital and dropping out of high school: Benefits to at-risk students of teachers' support and guidance. *Teachers College Record, 103*(4), 548-581.

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53-72.

DeHoon, M. J. L., Imoto, S., Nolan, J., & Miyano, S. (2004). Open source clustering software. *Bioinformatics, 20*(9), 1453-1454.

Delgado-Gaitan, C. (1988). The value of conformity: Learning to stay in school. *Anthropology & Education Quarterly, 19*(4), 354-381.

Demie, F. (2002). Pupil mobility and educational achievement in schools: An empirical analysis. *Educational Research, 44*(2), 197-215.

Demie, F., Lewis, K., & Taplin, A. (2005). Pupil mobility in schools and implications for raising achievement. *Educational Studies, 31*(2), 131-147.

Detert, J. R., Kopel, M. B., Mauriel, J., & Jenni, R. (2000). Quality management in U.S. high schools: Evidence from the field. *Journal of School Leadership, 10*(2), 158-187.

Dolled-Filhart, M., Ryden, L., Cregger, M., Jirstrom, K., Harigopal, M., Camp, R. L., et al. (2006). Classification of breast cancer using genetic algorithms and tissue microarrays. *Clinical Cancer Research, 12*, 6459-6458.

Dynarski, M., & Gleason, P. (2002). How can we help? What we have learned from recent federal dropout prevention evaluations. *Journal of Education for Students Placed at Risk, 2002*(1), 43-69.

Earle, L., & Fullan, M. (2003). Using data in leadership for learning. *Cambridge Journal of Education, 33*(3), 383-394.

Earle, L., & Katz, S. (2003). Leading schools in a data-rich world. In K. Leithwood & P. Hallinger (Eds.), *Second international handbook of educational leadership and administration* (pp. 1003-1022). Dordrecht, Netherlands: Kluwer Academic.

Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership, 31*(1), 15-24.

Ein-Dor, L., Zuk, O., & Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences, 103*(15), 5923-5928.

Eisen, M. B. (1998). Eisenlab. 2005, from http://rana.lbl.gov/EisenPublications.htm

Eisen, M. B., & DeHoon, M. (2002). *Cluster 3.0 manual*. Palo Alto, CA: Stanford University.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences, 95*, 14863-14868.

Elashoff, J. D., & Snow, R. E. (1971). *Pygmalion reconsidered*. Worthington, OH: Charles A. Jones Publishing Company.

Elmore, R. F. (2002). Leadership of instructional improvement. In *School reform and the superintendency: The 4th and 5th journals of the northeast superintendent's annual leadership institute* (Vol. 4, pp. 83-100). Providence: The LAB at Brown University.

Elmore, R. F. (2003). Accountability and capacity. In M. Carnoy, R. Elmore & L. S. Sisken (Eds.), *The new accountability: High schools and high stakes testing* (pp. 195-209). New York: RoutledgeFalmer.

Elmore, R. F., & Burney, D. (1999). Investing in teacher learning: Staff development and instructional improvement. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 263-291). San Francisco: Jossey-Bass.

Ensminger, M. E., & Slusarcick, A. L. (1992). Paths to high school graduation or dropout: A longitudinal study of a first-grade cohort. *Sociology of Education, 65*(2), 91-113.

Evans, F. B. (1976). What research says about grading. In S. B. Simon & J. A. Bellanca (Eds.), *Degrading the grading myths: A primer of alternatives to grades and marks* (pp. 30-50). Washington D.C.: Association for Supervision and Curriculum Development.

Falk, B. (2002). Standards-based reforms: Problems and possibilities. *Phi Delta Kappan, 83*(8), 612-620.

Farr, B. P. (2000). Grading practices: An overview of the issues. In E. Trumbull & B. Farr (Eds.), *Grading and reporting student progress in an age of standards* (pp. 1-22). Norwood: Christopher-Gordon Publishers.

Flanagan, K. S., Bierman, K. L., & Kam, C.-M. (2003). Identifying at-risk children at school entry: The usefulness of multibehavioral problem profiles. *Journal of Clinical Child and Adolescent Psychology, 32*(3), 396-407.

Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice, 12*(3), 23-30.

Fullan, M. (2000). The three stories of education reform. *Phi Delta Kappan, 81*(8), 581-585.

Fullan, M. (2001). *Leading in a culture of change*. San Francisco: Jossey-Bass.

Gamoran, A., & Hannigan, E. C. (2000). Algebra for everyone? Benefits of college-prepartory mathematics for students with diverse abilities in early elementary school. *Educational Evaluation and Policy Analysis, 22*(3), 241-254.

Girotto, J. R., & Peterson, P. E. (1999). Do hard courses and good grades enhance cognitive skills? In S. E. Mayer & P. E. Peterson (Eds.), *Earning and learning: How schools matter*. Washington DC: Brookings Institution Press.

Gleason, P., & Dynarksi, M. (2002). Do we know whom to serve? Issues in using risk factors to identify dropouts. *Journal of Education for Students Placed at Risk, 7*(1), 25-41.

Goslin, D. A. (1968). Standarized ability tests and testing. *Science, 159*(3817), 851-855.

Greene, J. P., & Caire, K. (2001). High school graduation rates in the United States. Retrieved July 6, 2006, from www.manhattan-institute.org/pdf/cr_baeo.pdf

Greene, J. P., & Winters, M. A. (2005). Public high school graduation and college-readiness rates: 1991-2002. Retrieved July 7, 2006, from www.manhattan-institute.org/html/ewp_081.htm

Guide to using data in school improvement efforts: A compilation of knowledge from data retreats and data use at learning point associates. (2004).). Naperville: Learning Point Associates.

Gutman, L. M., Sameroff, A. J., & Cole, R. (2003). Academic growth curve trajectories from 1st grade to 12th grade: Effects of multiple social risk factors and preschool child factors. *Developmental Psychology, 39*(4), 777-790.

Halverson, R. R., Prichett, R., Grigg, J., & Thomas, C. (2005). *The new instructional leadership: Creating data-driven instructional systems in schools* (No. WCER Working Paper No. 2005-9). Madison: University of Wisconsin-Madison, Wisconsin Center for Education Research.

Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eight grade. *Child Development, 72*(2), 625-638.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure. *Child Development, 76*(5), 949-967.

Hargis, C. H. (1990). *Grades and grading practices: Obstacles to improving education and helping at-risk students*. Springfield: Charles C. Thomas.

Hayman, J., Rayder, N., Stenner, A. J., & Madely, D. L. (1979). On aggregation, generalization, and utility in educational evaluation. *Educational Evaluation and Policy Analysis, 1*(4), 31-39.

Hightower, A. M., & McLaughlin, M. W. (2005). Building and sustaining an infrastructure for learning. In F. M. Hess (Ed.), *Urban school reform: Lessons from San Diego* (pp. 71-92). Cambridge, Mass.: Harvard Education Press.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall.

Jencks, C., & Phillips, M. (1999). Aptitude or achievement: Why do test scores predict educational attainment and earnings? In S. E. Mayer & P. E. Peterson (Eds.), *Earning and learning: How schools matter*. Washington DC: Brookings Institution Press.

Jimerson, S. R., Anderson, G. E., & Whipple, A. D. (2002). Wining the battle and losing the war: Examining the relation between grade retention and dropping out of high school. *Psychology in the schools, 39*(4), 441-457.

Jimerson, S. R., Egeland, B., Sroufe, L. A., & Carlson, B. (2000). A prospective longitudinal study of high school dropouts examining multiple predictors across development. *Journal of School Psychology, 38*(6), 525-549.

Jimerson, S. R., Pletcher, S. M. W., Graydon, K., Schnurr, B. L., Nickerson, A. B., & Kundert, D. K. (2005). Beyond grade retention and social promotion: Promoting the social and academic competence of students. *Psychology in the schools, 43*(1), 85-97.

Kallioniemi, A. (2002). Molecular signatures of breast cancer - predicting the future. *New England Journal of Medicine, 347*(25), 2067-2068.

Kamii, C., & Joseph, L. (2003). *Young children continue to reinvent arithmetic -- 2nd grade: Implications of Piaget's theory*. New York: Teachers College Press.

Kerr, K. A., Marsh, J. A., Schuyler-Ikemoto, G., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education, 112*(4).

Kienzi, G., & Kena, G. (2006). *Economic outcomes of high school completers and noncompleters 8 years later* (No. NCES 2007-019). Washington D.C.: National Center for Education Statistics.

Kirschenbaum, H., Napier, R., & Simon, S. B. (1971). *Wad-ja-get? The grading game in American education*. New York City: Hart Publishing Company.

Knesting, K., & Waldron, N. (2006). Willing to play the game: How at-risk students persist in school. *Psychology in the schools, 43*(5), 599-611.

Kohn, A. (1994). Grading: The issue is not how but why. *Educational Leadership, 52*(2), 38-41.

Kohonen, T. (1997). *Self-organizing maps*. New York: Springer.

Kuncel, N. R., Crede, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75*(1), 63-83.

Laird, J., DeBell, M., & Chapman, C. (2006). *Dropout rates in the United States: 2004*. Washington, D.C.: National Center for Education Statistics.

Langdon, H. W., & Trumbull, E. (2000). Grading and special populations. In E. Trumbull & B. Farr (Eds.), *Grading and reporting student progress in an age of standards* (pp. 129-156). Norwood: Christopher-Gordon Publishers.

Lehr, C. A., Hansen, A., Sinclair, M. F., & Christenson, S. L. (2003). Moving beyond dropout towards school completion: An integrative review of data-based interventions. *School Psychology Review, 32*(3), 342-364.

Linn, R. L. (1982). Ability testing: Individual differences, prediction, and differential prediction. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies* (pp. 335-388). Washington DC: National Academy Press.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Lorr, M. (1983). *Cluster analysis for social scientists: Techniques for analyzing and simplifying complex blocks of data*. San Francisco: Jossey-Bass, Inc.

Lortie, D. C. (1975). *Schoolteacher: A sociological study*. Chicago: University of Chicago Press.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saaverda, E., Lamb, J., Peck, D., et al. (2005). Mirco-RNA expression profiles classify human cancers. *Nature, 435*(9), 834-838.

Machin, S., Telhaj, S., & Wilson, J. (2006). The mobility of English school children. *Fiscal Studies, 27*(3), 253-280.

Madey, D. L. (1982). Some benefits of integrating qualitative and quantitative methods in program evaluation, with illustrations. *Educational Evaluation and Policy Analysis, 4*(2), 223-236.

Marrow, G. (1986). Standardizing practice in analysis of school dropouts. *Teachers College Record, 87*(3), 342-355.

Marshall, J. C. (1997). Data-based decision making. In S. D. Caldwell (Ed.), *Professional development in learning-centered schools* (pp. 150-167). Oxford, Ohio: National Staff Development Council.

Massell, D., & Goertz, M. E. (2002). District strategies for building instructional capacity. In A. M. Hightower, M. S. Knapp, J. A. Marsh & M. W. McLaughlin (Eds.), *School district and instructional renewal* (pp. 43-60). New York: Teachers College Press.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Meyer, R. H. (1999). The effects of math and math-related courses in high school. In S. E. Mayer & P. E. Peterson (Eds.), *Earning and learning: How schools matter*. Washington DC: Brookings Institution Press.

Militello, M. (2004). *At the cliff's edge: Utilizing evidence of student achievement for instructional improvement in a school district.* Michigan State University, E. Lansing Michigan.

Miller, M. J. (2005). *Accounting for student and school success in high-poverty, high-minority schools: A constructivist approach.* Unpublished Dissertation, University of Texas San Antonio, San Antonio, Texas.

Montes, G., & Lehmann, C. (2004). *Who will drop out from school? Key predictors from the literature* (No. T04-001). Rochester, NY: Childrens Institutue Inc.

Murphy, J., & Hallinger, P. (2001). Characteristics of instructionally effective school districts. *Journal of Educational Research, 81*(3), 175-181.

NCES. (2001). International comparisons in fourth-grade reading literacy: Findings from the progress in international reading literacy study (PIRLS) of 2001. Retrieved February 16, 2007, from http://nces.ed.gov

NCES. (2004). National institute of statistical sciences/education statistics services institute task force on graduation, completion, and dropout indicators. In U.S. Dept. of Education (Ed.): National Center for Education Statistics.

NCES. (2006). Common core of data. Retrieved December 18, 2006, from http://nces.ed.gov/ccd/

Newmann, F. M. (1991). Linking restructuring to authentic student achievement. *Phi Delta Kappan, 72*(6), 458-463.

Onwuegbuzie, A. J., & Leech, N. L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology, 8*(5), 375-387.

Parsons, T. (1959). The school class as a social system: Some of its functions in American society. *Harvard Educational Review, 29*(4), 297-318.

Popham, J. W. (2004). "Teaching to the test": An expression to eliminate. *Educational Leadership, 62*(3), 82-83.

Porter, A. C., & Smithson, J. L. (2001). Are content standards being implemented in the classroom? A methodology and some tenative answers. In S. H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states*. Chicago: The University of Chicago Press.

Quann, C. J. (1983). *Grades and grading: Historical perspectives and the 1982 AACRAO study*: American Association of Collegiate Registrars and Admissions Officers.

Randolph, K. A., & Orthner, D. K. (2006). A strategy for assessing the impact of time-varying family risk factors on high school dropout. *Journal of Family Issues, 27*(7), 933-950.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology, 76*(85-97).

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher, 34*(5), 25-31.

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

Rencher, A. C. (2002). *Methods in multivariate analysis* (2nd ed.). Hoboken: John Wiley & Sons, Inc.

Roderick, M., & Camburn, E. (1999). Risk and recovery from course failure in the early years of high school. *American Educational Research Journal, 36*(2), 303-343.

Roderick, M., Nagaoka, J., Bacon, J., & Easton, J. Q. (2000). Update: Ending social promotion. Retrieved June 21, 2006, from http://www.consortium-chicago.org/publications/pdfs/p0g01.pdf

Romesburg, H. C. (1984). *Cluster analysis for researchers*. Belmont, CA: Lifetime Learning Publications.

Rosenthal, R., & Jacobsen, L. (1969). *Pygmalion in the classroom: Self fulfilling prophecies and teacher expectations*. New York: Holt, Rinehart and Winston.

Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform to close the black-white achievement gap*. Washington, DC: Economic Policy Institute.

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal, 32*(3), 583-625.

S&P. (2006). School matters. Retrieved December 18, 2006, from http://www.schoolmatters.com/

Salpeter, J. (2004). Data: Mining with a mission. *Technology & Learning, 24*(8), 30-32,34,36.

Schmoker, M. (1999). *Results: The key to continuous school improvement* (2nd ed.). Alexandria, Virgina: Association for Supervision and Curriculum Development.

Seastrom, M., Hoffman, L., Chapman, C., & Stillwell, R. (2006). The average freshman graduation rate for public high schools from the common core of data: School years 2002-03 and 2003-04. In U. S. Dept. of Education (Ed.): Institute of Education Sciences.

Secada, W. G. (2001). From the director: Using data for educational decision-making. *The Newsletter of the Comprehensive Center-Region VI, 6*(1), 1-2.

Shaffer, D. W., & Serlin, R. C. (2004). What good are statistics that don't generalize? *Educational Researcher, 33*(9), 14-25.

Shen, R., Ghosh, D., Chinnaiyan, A., & Meng, Z. (2006). Eigengene-based linear discriminant model for tumor classification using gene expression microarray data. *Bioinformatics, 22*(21), 2635-2642.

Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F., Snowden, J. B., Gordon, E., et al. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do*. San Francisco: Jossey-Bass.

Sima, C., & Dougherty, E. R. (2006). What should be expected from feature selection in small-sample settings. *Bioinformatics, 22*(19), 2430-2436.

Simon, S. (1976). An overview. In S. B. Simon & J. A. Bellanca (Eds.), *Degrading the grading myths: A primer of alternatives to grades and marks* (pp. 1-4). Washington D.C.: Association for Supervision and Curriculum Development.

Simon, S. B., & Bellanca, J. A. (1976). *Degrading the grading myths: A primer of alternatives to grades and marks*. Washington DC: Association for Supervision and Curriculum Development.

Sipple, J. W., Killeen, K., & Monk, D. H. (2004). Adoption and adaptation: School district responses to state imposed learning and graduation requirements. *Educational Evaluation and Policy Analysis, 26*(2), 143-168.

Sireci, S. G., Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitiate standard setting. *Applied Measurement in Education*.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification*. San Francisco: W.H. Freeman.

Sorlie, T., Perou, C. M., Fan, C., Geisler, S., Aas, T., Nobel, A., et al. (2006). Gene expression profiles do not consistently predict the clinical treatment response in locally advanced breast cancer. *Molecular Cancer Therapy, 5*, 2914-2918.

Spitz, H. H. (1999). Beleaguered pygmalion: A history of the controversy over claims that teacher expectancy raises intelligence. *Intelligence, 27*(3), 199-234.

SPSS. (2006). SPSS. Retrieved December 18, 2006, from http://www.spss.com/

Starch, D., & Elliot, E. C. (1912). Reliability of grading of high school work in English. *School Review, 21*, 442-457.

Starch, D., & Elliot, E. C. (1913a). Reliability of grading work in mathematics. *School Review, 22*, 254-259.

Starch, D., & Elliot, E. C. (1913b). Reliability of the grading of high school work in history. *School Review, 21*, 676-681.

Strand, S., & Demie, F. (2006). Pupil mobility, attainment and progress in primary school. *British Educational Research Journal, 32*(4), 551-568.

Streifer, P. A. (2002). *Using data to make better educational decisions*. Lanham, Maryland: The Scarecrow Press with the American Association of School Administrators.

Streifer, P. A. (2004). *Tools and techniques for effective data-driven decision making*. Lanham: Scarecrow Education.

Streifer, P. A. (2005). Using data mining to identify actionable information: Breaking new ground in data-driven decision making. *Journal of Education for Students Placed at Risk, 10*(3), 281-293.

Supovitz, J. A. (2002). Developing communities of instructional practice. *Teachers College Record, 104*(8), 1591-1626.

Swanson, C. B. (2004, Feb). Who graduates? Who doesn't? A statistical portrait of public high school graduation, class of 2001. Retrieved July 7, 2006, from www.urban.org/UploadedPDF/410934_WhoGraduates.pdf

Teddlie, C. (1994). The integration of classroom and school process data in school effectiveness research. In D. Reynolds, B. P. M. Creemers, P. S. Nesselrodt, E. C. Schaffer, S. Stringfield & C. Teddlie (Eds.), *Advances in school effectiveness research and practice* (pp. 111-132). Tarrytown NY: Elsevier Science.

Thorn, C., A. (2002). *Data use in the school and classroom: The challenges of implementing data-based decision making inside schools*. Madison: University of Wisconsin-Madison, Wisconsin Center for Education Research.

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology, 40*, 7-26.

Trumbull, E. (2000a). Avoiding bias in grading systems. In E. Trumbull & B. Farr (Eds.), *Grading and reporting student progress in an age of standards* (pp. 105-128). Norwood: Christopher-Gordon Publishers.

Trumbull, E. (2000b). Why do we grade - and should we? In E. Trumbull & B. Farr (Eds.), *Grading and reporting student progress in an age of standards* (pp. 23-44). Norwood: Christopher-Gordon Publishers.

Trusty, J. (2002). Effects of high school course-taking and other variables on choice in science and mathematics collge majors. *Journal of Counseling and Development, 80*(4), 464-475.

Tyler, J. H. (2003). Economic benefits of the GED: Lessons from recent research. *Review of Educational Research, 73*(3), 369-403.

U.S. Census bureau. (2007). Retrieved March, 16, 2007, from www.census.gov

van'tVeer, L. J., Dai, H., vandeVijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*, 530-536.

vandeVijver, M. J., He, Y. D., van'tVeer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine, 347*(25), 1999-2009.

Viadero, D. (2006). Signs of early exit for dropouts abound. *Education Week, 25*(41S), 20-22.

Vilo, J. (2003). Expression profiler - epclust. 2005, from http://ep.ebi.ac.uk/EP/EPCLUST/

Waters, L. B. (2000). How to design a model standards-based accountability system. In E. Trumbull & B. Farr (Eds.), *Grading and reporting student progress in an age of standards* (pp. 87-103). Norwood: Christopher-Gordon Publishers.

Wayman, J. C. (2005). Involving teachers in data-driven decision making: Using computer data systems to support teacher inquiry and reflection. *Journal of Education for Students Placed at Risk, 10*(3), 295-308.

Wayman, J. C., & Stringfield, S. (2006a). Data use for school improvement: School practices and research perspectives. *American Journal of Education, 112*(4), 463-468.

Wayman, J. C., & Stringfield, S. (2006b). Technology-supported involvement of entire faculties in examination of student data for instructional improvement. *American Journal of Education, 112*(4), 549-571.

Wayman, J. C., Stringfield, S., & Yakimowski, M. (2004). *Software enabling school improvement through the analysis of student data (report no. 67)*. Baltimore, MD: Johns Hopkins University, Center for Researc on the Education of Students Placed At Risk.

Weber, J. M. (1989). *Identifying potential dropouts: A compliation and evaluation of selected procedures*. Columbus, OH: Ohio State University.

Weinstein, J. N., Myers, T. G., O'Connor, P. M., Friend, S. H., Fornace, A. J., Kohn, K. W., et al. (1997). An information-intensive approach to the molecular pharmacology of cancer. *Science, 275*(5298), 343-349.

Wells, C. A. (2003). Providing highly mobile students with effective education. Eric digest. from http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/2a/3b/7a.pdf

Wightman, L. F. (1993). *Clustering United States law schools using variables that describe size, cost, selectivity, and student body charateristics*. Newtown, PA: Law School Admission Council.

Wood, L. A. (1994). An unintended impact of one grading practice. *Urban Education, 29*(2), 188-201.

Woodruff, D. J., & Ziomek, R. L. (2004). *High school grade inflation from 1991 to 2003. Research report series 2004-04*. Iowa City, IA: ACT, Inc.

Woods, E. G. (1995). *Reducing the dropout rate* (No. Close-Up #17): Northwest Regional Educational Laboratory.

Yore, L. D., Bisanz, G. L., & Hand, B. M. (2003). Examing the literacy component of science literacy: 25 years of language arts and science research. *International Journal of Science Education, 25*(6), 689-725.

Young, S., & Shaw, D. G. (1999). Profiles of effective college and university teachers. *The Journal of Higher Education, 70*(6), 670-686.

Young, V. M. (2006). Teachers' use of data: Loose coupling, agenda setting and team norms. *American Journal of Education, 112*(4), 521-548.

Zapala, M. A., & Schork, N. J. (2006). Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the National Academy of Sciences, 103*(51), 19430-19435.

Zvoch, K. (2006). Freshman year dropouts: Interactions between student and school characteristics and student dropout status. *Journal of Education for Students Placed at Risk, 11*(1), 97-117.