2020

Case Study: Formative Assessment Driven Instruction and Standardized Test Scores for 10th Grade Biology Students

Gregory Laurence Nelson
Bethel University

Case Study: Formative Assessment Driven Instruction and Standardized Test Scores for 10th Grade Biology Students

by

Gregory Laurence Nelson

A dissertation submitted to the faculty of Bethel University in partial fulfillment of the requirements for the degree of Doctor of Education.

St. Paul, MN

2020

Approved by:
Dr. Aldo Sicoli, adviser
Dr. Jennifer Hill, reader
Dr. Craig Paulson, reader

Abstract

Accountability measures have been employed in United States schools to meet the demands of a globalized society with a standardized testing system used to assess student growth. The purpose of this study was to determine the relationship between a research-backed pedagogical instructional approach, formative assessment-driven instruction, and success in a large-scale standardized test system. Standardized testing is a practical necessity for an accountability system and if an authentic instructional process could be of support, a key piece of evidence will be brought forward to the educational equation. This was a quantitative ex-post facto archival case study. An analysis of data over the five years of this study showed no significant relationship between a formative assessment-driven instructional approach and improved standardized test scores.

Acknowledgements

Thank you Janet for being my soul mate for over forty years. You are an amazing person and the brightest educator I know. I am deeply grateful for the love and support from my entire family throughout this process. Mom and dad: I wish you could be here for this. I miss you both.

Table of Contents

List of Tables……8

List of Figures……...9

Chapter I: Introduction……10

Introduction to the Problem……..10

Background of the Study……..10

Statement of the Problem……..19

Purpose of the Study…….22

Research Question…….22

Hypothesis…….23

Significance of the Study……...24

Rationale……24

Definition of Terms……25

Assumptions and Limitations…….25

Nature of the Study…….26

Organization of the Remainder of the Study……..26

Chapter II: Literature Review……28

Introduction……....….28

Formative Assessment Definitions…….....28

Formative Assessment Components……...29

Formative Assessment Student – Teacher Interactions……..31

Formative Assessment Theoretical Foundations……33

Formative Assessment Grading Considerations……..34
Standardized Testing Overview……...……36
Standardized Testing Complexities……..37
Standardized Testing Criticism……38
Standardized Testing Proponents…….39
Standardized Testing Results……...40
Summary……..42
Chapter III: Methodology……...43
Introduction…….....…43
Purpose of Study…….....43
Conceptual Framework……...45
Research Design……. 45
Research Question……...46
Hypotheses……..46
Variables……. 47
Instruments and Measures…….. 47
Sampling Design……..... 48
Data Collection Procedures……...…… 50
Data Analysis……. 51
Reliability, Validity, and Trustworthiness……. 51
Limitations and Assumptions…… 52
Ethical Considerations……... 53

Chapter IV: Results…… 55

Overview…….55

2011…….55

2012…….60

2013…… ……65

2014…….....70

2015…….75

Summary of Results…….……80

Chapter V: Summary……...81

Introduction……..81

Overview of Study……...81

Research Question……...82

Hypotheses……..82

Analysis……...…83

Conclusion……...84

References…….. 87

List of Tables

1. 2011 Means and Standard Deviations for MCA-II Scores by Class Type and GPA Quartiles……..56

2. 2 (Class) X 4 (GPA Quartiles) ANOVA Table for 2011 MCA-II Scores…….....57

3. Tukey's HSD Post Hoc Tests for 2011 MCA-II Scores by GPA Quartiles……...58

4. 2012 Means and Standard Deviations for MCA-II Scores by Class Type and GPA Quartiles……..61

5. 2 (Class) X 4 (GPA Quartiles) ANOVA Table for 2012 MCA-II Scores …….....62

6. Tukey's HSD Post Hoc Tests for 2012 MCA-II Scores by GPA Quartiles……...63

7. 2013 Means and Standard Deviations for MCA-II Scores by Class Type and GPA Quartiles……..66

8. 2 (Class) X 4 (GPA Quartiles) ANOVA Table for 2013 MCA-II Scores……..67

9. Tukey's HSD Post Hoc Tests for 2013 MCA-II Scores by GPA Quartiles……68

10. 2014 Means and Standard Deviations for MCA-III Scores by Class Type and GPA Quartiles……..71

11. 2 (Class) X 4 (GPA Quartiles) ANOVA Table for 2014 MCA-III Scores……...72

12. Tukey's HSD Post Hoc Tests for 2014 MCA-III Scores by GPA Quartiles…….73

13. 2015 Means and Standard Deviations for MCA-III Scores by Class Type and GPA Quartiles……...76

14: 2 (Class) X 4 (GPA Quartiles) ANOVA Table for 2015 MCA-III Scores……77

15: Tukey's HSD Post Hoc Tests for 2015 MCA-III Scores by GPA Quartiles…….....78

List of Figures

1. 2011 MCA-II Scores by Class Type and GPA Quartiles……..59

2. 2012 MCA-II Scores by Class Type and GPA Quartiles…….....64

3. 2013 MCA-II Scores by Class Type and GPA Quartiles……..69

4. 2014 MCA-III Scores by Class Type and GPA Quartiles…….……74

5. 2015 MCA-III Scores by Class Type and GPA Quartiles……..…79

Chapter I: Introduction

Introduction to the Problem

The nations of the world have a commonality among them in that they include an educational process as a vital component of their systems and structures. The system is used to train new members of society and provide support for geopolitical, social, and economic success and relevance. The United States has moved through its historical eras with its education system developing and adapting along the way to meet the evolving geopolitical, social, and economic needs of the day. The current twenty-first century globalized era places complex demands on its system as never before. A substantial retooling is required for the broad purpose of producing a more widely educated populace. Current and emerging economic realities have literally restructured the makeup of the workplace.

The essential question for this retooling has been how the United States should restructure its education system for maximum effectiveness for the changing realities of the twenty first century globalized world. The process has yet to achieve a successful result. Reform attempts have either folded under political pressure or been difficult to implement on a large scale. In the midst of bulk reform efforts, promising classroom level pedagogy research has emerged that may provide a bridge towards the elusive results needed in the global era.

Background of the Study

The fading of the Cold War in the latter 1980s and early 1990s produced the emerging global era. The bi-polar equation that had ordered the world for decades fell apart. Thomas Friedman (2005) became an early and prominent analyst in capturing exactly what was happening. "Disrupting forces were unleashed altering or even eliminating traditional structures in commerce, labor, government, communication, and travel" (p. 182). He went on to coin the

notion that the world was now "flat." The new era was described as having "new players, on a new playing field, developing new processes and habits for horizontal collaboration" (p.182). He further specified that "technological advances in communication, travel, and automation have compressed the world in terms of time and space producing a hyper paced world economy" (p. 182).

These new global realities produced an entirely new equation for the educational system to solve. Significant changes were in order. Futurist Alvin Toffler (1990) had previously framed the challenge that presented itself when he wrote; "the illiterate of the twenty first century will not be those who cannot read and write but those who cannot learn, unlearn, and relearn" (p. 2). Knowledge had become the crux of world power. Education was charged with the task of reinventing itself to meet the needs of this emerging era.

At the core of the emerging system was a need for an "across the board" educated workforce. Many entry level jobs now had requirements for reading levels higher than those previously required for high school graduation. Automation steadily eliminated a significant number of low skill jobs and added ones that required significant problem-solving skills. Emerging technology had created whole new economic sectors with each of these accompanied by the need for highly skilled and educated workers. The pivotal 1990 report by the National Center on Education and the Economy entitled *America's Choice: High Skills or Low Wages* underscored those changes. High levels of educational achievement in the globalized world were no longer exclusive to managerial levels. The educational implications for the change were significant. "In essence, society moved the goalposts" (Christensen, 2008, p. 58).

The democratic foundations of the United States, over time, had created a demand for increased access to education. Once a privilege afforded only to the wealthy, it expanded over

time to include a much broader swath of citizenry eventually morphing into a public and free system for all. This was largely accomplished by the mid-twentieth century. It helped fuel the maturation of the United States into an industrialized world power that would triumph in two world wars (Christensen, 2008).

The "total war" effort of the United States during World War II laid the groundwork for major societal change when Black Americans served in the armed forces. There would be no "turning back the clock" on this significant dent into the segregation systems in place in the United States. A Civil Rights movement emerged and among the measures that drove it was the idea that the free and public-school system of the United States should benefit all citizens, regardless of race and ethnicity. In the years to follow, the desegregation of schools, particularly in southern states, dominated educational headlines. The public education system of the United States had gained a new responsibility: to be an institution that would produce a "level playing field" for all.

The end of World War II had brought the United States into full superpower status and an intense Cold War competition with its former ally, the Soviet Union. Cold War dynamics had produced a need to compete technologically with the "other side." The space and arms races with the Soviet Union resulted in ramped up attention to the teaching of mathematics and science. But economic forces as a whole were quite content to let the education system pass on a sorted mixed product. There was room in the economy for low skilled and educated workers. The traditional white- and blue-collar categories took it from there. School was generally seen as a place that provided opportunity. It was up to the student to take advantage of it (Christensen, 2008).

The effort to retool American education for the globalized twenty-first century can be traced to 1981 when the U.S. Secretary of Education created the National Commission of Excellence in Education to address competitive concerns. The newly created organization released the landmark study, *A Nation at Risk*, two years later. The study indicated that a mature United States had lost its competitive edge in the then rapidly globalizing world and a call for reform was sounded. The report became a pivotal moment in United States education history providing a basis for future school choice, testing, and accountability initiatives. Subsequent studies contradicted the report showing steady or slightly improving test scores during the same time period. Two of the original authors indicated the report was not an objective study but intended to send an alarm bell (Kamenetz, 2018).

The prevailing view that emerged from the study was that the global dominance that came as a result of being the victor in multiple world wars was fading. Technical innovations elsewhere began challenging U.S. companies. "Public confidence in schools began sagging, especially when compared to the 1940s and 1950s, and the nation asked its schools to take on the job of keeping the United States competitive" (Christensen, 2008, p. 58).

The release of *A Nation at Risk* (1983) had a long-lasting effect that guided the educational narrative for nearly four decades. Schools were identified as the source of the problem in education and also as the solution. Societal factors, traditionally a key component in education reform conversations, were included in the report but given far less billing in the work. Schools were to be responsible for student outcomes. The emphasis on school responsibility created a fundamental divide that has persisted through subsequent battles over school reform, with many teachers arguing that it is unfair for them to be judged on outcomes that are at least partly out of their control, and with political reformers preaching accountability (Mehta, 2013).

The reform movement, post *A Nation at Risk,* was tied to the political climate of the day. The ending of the Cold War had convinced many educational reform was realistic. The potential benefits of testing, accountability, choice and markets were viewed as entities that could guarantee that tax dollars invested in education were getting a good return. Accountability became the watchword of public officials and business leaders (Ravitch, 2010).

The accountability concept took shape when it was moved forward by state governors in 1986. The governors had become very concerned about the jobs that were being lost to low-wage countries, and business leaders began to realize that skilled and educated people were vital to their future. The governors took the initiative and outlined a general approach to guide reforming the educational process based on goal setting coupled with an efficient process to achieve them. The process of monitoring goals work, standards, would create irresistible pressure on the schools to find effective curriculum materials, implement effective instructional strategies, and do the other things needed to raise student performance (Tucker & Codding, 2002).

The efforts of the governors brought the standards movement to the federal level where they quickly became caught up in political crossfire. The standards conversation, in the eyes of many reformers, would produce a national entity that would in turn create some sort of national system of standards and assessments. Many assumed a national test, similar to most European and Asian nations, would be the ultimate result. Republicans and Democrats took turns at various formulations of ways in which the federal government could take the lead in creating such a system. At each step along the way, specific proposals stirred political concern. Political conservatives in general feared that entrusting these functions to the federal government could lead to the imposition of a national curriculum. Such a curriculum would be used by the "other

side" to erode traditional values entrenching much of the 1960's social movement into the mainstream. Liberals, on the other hand, feared that the lack of national cohesion would harm poor and minority students because of the inequitable distribution of resources in the American education system (Tucker & Codding, 2002).

A "beginning of the decade" goal to establish a system of standards enabling the United States to graduate students with an education comparable to that offered by any nation in the world had not come to fruition. States found it difficult to realign the moving parts of education. The numerous stakeholders involved: textbook publishers, test publishers, and schools of education produced insurmountable political obstacles. The standards that some states did adopt were lacking as political processes produced standards that were completely non-controversial (Ravitch 2010). By 2001, all initiatives to create national exams or tests, to reference state tests to national tests, and to review state standards and tests at the national level had failed.

The second vein of reform, accompanying accountability, was the area of choice. Conceptually, parents would be given options within the umbrella of the public-school system driven by innovation per market principles. Competition in turn would spur public schools to mirror the success of the option schools. Slow to change public schools could no longer maintain ineffective systems if students legitimately could take their business, including funding, elsewhere. Choice options eventually included open enrollment (ability to enroll in a public system regardless of residence), various voucher systems (government issued educational coupons), charter schools (public school operated according to charter as opposed to state regulations), and postsecondary enrollment (ability for high school students to take college courses). The advent and development of online instructional delivery also enhanced choice options over time.

The choice movement made clear inroads into the United States public school system as a whole. The U.S. Department of Education (2019) reported a decline in students attending their local "assigned" public schools from 74% in 1999 to 69% in 2016. Those attending "choice" public schools increased 5% over the same time period to 19%. Homeschool students rose slightly from 2% to 3% while private school enrollment fell from 11% to 9%. Thirteen different states plus the cities of Cleveland and Milwaukee operated voucher programs. Five of those 13 states plus 11 others implemented tax credit scholarships and/or personal tax credit programs. Yet, while these significant choice options have become a reality, the movement has produced inconclusive student achievement results.

With the standards movement ineffectiveness and choice options having nominal impact, school reform tacked in a different direction. Elected officials became convinced that a system of measurement and data would fix schools. An accounting strategy emerged with rewards and consequences attached. Standardized tests would be developed and deployed as units of measurement. In short order, school accountability became synonymous with standardized test results (Ravitch, 2010).

This directional shift emerged in full force following the presidential election of 2000. The signature legislative milestone of the test results directional change was dubbed the No Child Left Behind Act (NCLB) of 2001. NCLB was actually the reauthorization of the previously passed Elementary and Secondary Education Act of 1965. Passed by Congress in 2001 with clear bipartisan support, NCLB was signed into law by President George W. Bush in January of 2002 and remained intact until 2015. The law greatly increased the federal government's role in education, especially in terms of holding schools accountable for the academic performance of their students. Although NCLB covered numerous federal education

programs, the law's requirements for testing, accountability, and school improvement received the most attention. NCLB required states to test students annually in both English language arts and mathematics in Grades 3-8, as well as once in Grades 10-12. States must also test students in science three times: once in the grade span of Grades 3-5, again in 6-8, and a final time in 10-12. Individual schools, school districts, and states were required to publicly report test results for all students, as well as for specific student subgroups, including low-income students, students with disabilities, English language learners, and major racial and ethnic groups. The goal was to level the playing field for disadvantaged students including those affected by poverty, students of color, and those receiving special education services.

NCLB measures produced a number of positive impacts. The measurement of student progress became an everyday reality for schools leading to greater inclusion. If all students are expected to achieve, all must be measured. The expectation was set that struggling students would learn alongside their peers, including those receiving special education services. Schools were pushed to give all students the attention, support and help they needed. Graduation rates showed improvement in the NCLB era moving from 57% in 2002 to 68% in 2011. Opportunity gap progress followed suit. The National Assessment of Educational Progress data show that the nation's minorities made substantial strides at Grades 4 and 8, especially in mathematics. In 1990, only one percent of Black 4th graders were proficient. By 2011, 17 % were. Hispanics went from five % proficient in 1990 to 24 % proficient in 2011. For both minorities, the gains in mathematics and reading between 1990 and 2011 in Grade 8 were only slightly less impressive (U.S. Department of Education, 2019).

It should be noted that there were other benefits to NCLB. States gained flexibility in how they spent federal funding, as long as schools were improving. Teachers were now required

to be highly qualified in the subjects they taught. Special education teachers had to be certified and demonstrate knowledge in every subject they teach. Finally, schools were required to employ research-based instructional methods (No Child Left Behind Act, 2001).

The NCLB movement began to lose momentum as it entered its second policy decade. While improvements did occur, they were considered modest at best. The standardized test results of the NCLB initiative did not meet the level playing field standard. Achievement gaps between white and minority student groups remained problematic. Concerns grew over the sheer numbers of failing schools (McNeil, 2011). The results were also problematic internationally as the United States sputtered in worldwide rankings (U.S. Department of Education, 2011).

The dependence of NCLB on standardized testing became problematic in the eyes of many. "Teaching to the test" became a focal point of curricular efforts leaving little time for other learning opportunities. Consequences for not meeting goals could be excessively harsh such as the firing of an entire school staff or even the closing of a struggling school. Critics linked several cheating scandals to NCLB, citing the pressure on teachers and educators to perform. Others argued that NCLB's standards-based accountability was inconsistent with special education, which focuses on meeting a child's individual needs (McNeil, 2011).

As NCLB enthusiasm waned, it was replaced with the *Every Student Succeeds Act* (ESSA) in December of 2015. The new act was a blend of old and new: some parts of NCLB were repealed with new features added. Requirements for highly qualified teachers, research-based instruction and basic reporting on school results were included in the new act. Standardized test scores remained a requirement of the new system with the closing of achievement gaps an expectation. Graduation rates received more intense focus, sharing the

stage with standardized test scores in addressing achievement gaps. The specifics of ESSA are currently being phased in on a state by state basis (Ferguson, 2016).

To date, the general consensus is that real educational progress has been elusive in the reform era. While national measures of student learning have generally inched forward, the results can be considered mediocre at best. Zip codes remain a strong predictor of student success. The broad sweeping systems changes of the schools fell short of intended outcomes. They did not produce results required by a globalized world (U.S. Department of Education, 2016).

Statement of the Problem

The tremendous investment of resources into education across nearly four decades of reform has produced rich pedagogical growth and clear blueprints for school effectiveness. Brain research has opened new instructional frontiers to new learning realities. Research established that high quality teaching would indeed produce better student achievement. Reform models such as Effective Schools, Accelerated Schools, and Schools Within Schools added valuable insight into effective schooling. The school reform movement was based on the premise that these revelations could be woven into a continuous improvement mindset for schools that would result in a far better product. The question of why this has not happened has left U.S. education analysts in a quandary (Wilburn, Cramer, & Walton, 2020).

Black and Wiliam (1998b), among others, noted along the way that reform work was falling short. "But the sum of all these reforms has not added up to an effective policy because something is missing" (p. 1). What was missing in their estimation, was a focus on the process of teaching and learning. Black and Wiliam (1998b), noted the work of Stiger and Hiebert (1999) in making the case for the absent ingredient noting that "a focus on standards and

accountability that ignores the processes of teaching and learning in classrooms will not provide the direction that teachers need in their quest to improve" (p. 1). They believed the educational reform movement had come to focus far too heavily on "outputs" given certain "inputs." Inputs (pupils, teachers, resources, rules, and requirements) were mixed together with assumed specified outputs (standardized test scores) expected to be produced. In their analysis, heavy emphasis was being focused on the outputs but little on the interaction of the inputs that were expected to produce specified outputs. The interaction of the inputs took place in the classroom with little oversight. The classroom, in effect, was treated like a black box (p. 1).

Black and Wiliam's extensive 1998 study, titled *Inside the Black Box: Raising Standards Through Classroom Assessment*, showed that the use of formative assessment as an instructional method showed significant results in student learning. The major premise of the study was threefold: evidence exists that the practice of formative assessment raises standards, there is room for improvement in the use of formative assessment, and there is evidence about how to improve formative assessment (p. 2). They challenged governments, their agencies, school authorities and the teaching profession to use the evidence for the purposes of raising standards in schools. They followed with a call for a paradigm shift: "we also acknowledge widespread evidence that fundamental change can be achieved only slowly - through programs of professional development that build on existing good practice" (p.2).

Black and Wiliam's (1998b) assertion that the elusive improved education product sought so intently during the reform decades would take place methodically over time flew in the face of the urgency that had characterized school reform efforts. Legislative measures typically promised results within specified election cycles. Lack of immediate success would provide ammunition for political opponents to take change in yet another direction. Automated testing

mechanisms produced streams of data that would in turn yield complex success analytics which would in turn suggest new goals. Moving education reform forward under a formative assessment banner would necessitate a pace not consistent with incompatible school reform realities of the global era. Yet, Black and Wiliam posited that a reversal of priorities where process becomes the focal point, would be successful. Research substantiated that changing instructional habits to a formative assessment approach would improve student performance.

A shift towards a formative assessment system would have inherent difficulty in the current mindset focused on short term success determinants. A gradual developmental process does not easily lend itself to statistical analysis. The open ended and varied nature of formative assessment is not a natural fit with large scale quantitative data collection processes. The practice of formative assessment typically manifests itself in learning activities that are flexible and adaptable to individual learner needs. But for a reform process to be deemed successful in the United States, evidence would need to be produced on a massive scale. The dominant collection mechanism in place is the standardized testing system (Koretz & Hamilton, 2006).

A formative assessment-driven system would produce a paradoxical duality. A research-backed approach with a strong likelihood of success was virtually unquantifiable on a large scale. The nationwide scope of the sweeping changes of the reform era had promoted a top down implementation system that focused on quantifiable results. Improved classroom instruction and improved student achievement were the desired outcomes of both the formative assessment-driven and the quantifiable standardized testing approaches. Investing needed resources and efforts to fully develop and widely implement a formative assessment-driven system will not come from an immediate dismantling of the current quantifiable standardized testing structure. Black and Wiliam (1998) acknowledged this stating that formative assessment-

driven instruction is not a “magic bullet” for education. “The issues involved are too complex and too closely linked to both the difficulties of classroom practice and the beliefs that drive public policy” (p.2).

Purpose of the Study

The purpose of this study was to determine whether or not a connection could be established between a decentralized formative assessment approach and success in a large-scale standardized test system. The former has significant research support while the latter is pragmatically necessary. The objective nature of a standardized testing system contradicts the individualization of the formative process. However, if an individualized formative process can indeed support a massive scale quantifiable outcome, the elusive journey to successful educational reform could be refocused in a researched-based optimistic manner.

Research Question

Research shows that better teachers produce better test scores from the students they teach. Marzano, Pickering, and Pollock (2001) found that properly implemented instructional strategies could result in percentile gains of 29–45 points in student achievement. Wright, Horn, and Sanders (1997) noted improving the effectiveness of teachers improved student achievement more than any other single factor. They further noted that effective teachers were effective with students of all achievement levels. Darling-Hammond (2000) and Stronge (2002) specified the ability to use a range of teaching strategies skillfully as a central characteristic of effective teachers. Goe and Stickler (2008) established a strong correlation between teacher quality and student achievement.

Yet, when the array of issues affecting student achievement were considered, it was apparent that some students are better positioned than others to learn in the current system.

When the effects of poverty, learning disabilities, and transiency were considered, a clear impact existed on student achievement results. Some students simply were better equipped than others to perform well in schools. The whole focus of the reform movement has been to widen the scope of school effectiveness to include all students.

The current standardized output structure of the education system favored those students who typically did well in school. The standardized test scores, by default, did show where student achievement is lacking. A standardized results system coupled with a specific learning process could provide evidence on how an educational process can truly serve a wider swath of students with a clearer path towards better achievement for all. The research question for this study was: what is the relationship between formative assessment-driven instruction and standardized test scores, particularly for average or below average students?

Hypothesis

The hypothesis of this study compared two groups of students who were assessed through a standardized testing event. The test was based on standards taught during two years of science courses. The standards were embedded in the two years of courses, with the vast majority in year two. One group of students was instructed in formative assessment-driven approach, the other in a traditional approach. The hypothesis proposed that students taught in a formative assessment-driven instructional approach would show better standardized test scores than students in the traditionally instructed group. The impact investigated was in relation to students' usual academic performance with a stronger relationship occurring in students with lower achievement levels.

Significance of the Study

The potential benefits for this study were significant. It was widely agreed that student achievement needs to improve in the United States given current global dynamics (Gordon, 2007). The World Economic Forum documented the decline of the United States educational output in The Global Competitiveness Report of 2016-2017 (Schwab, 2016). Further complicating the need is that a good share of this improved achievement needs to come from historically underperforming student groups. With a standardized testing system seemingly the only logical measuring system workable for school accountability, an instructional mindset that can build capacity for higher quality results would be a game changer. If formative assessment-driven teaching could be linked to improved test results, a worthy path to pursue accountability could be established. Whatever is done in the U.S. education system has to be done on a massive scale. Establishing a link between the two would provide an effective foundation from which true improvement would result. American education policy affects millions of students, families, teachers, and administrators. Establishing a link between a research proven method of instruction that needs sustained effort, attention, and resources with results that show improved 21st century compatible student learning could be invaluable for education direction.

Rationale

This study proposed that formative assessment-driven instruction had the potential to provide a bridge between effective classroom instruction and successful student achievement results from standardized tests. If an authentic instructional process, where assessment drives the pace and scope of learning, could have been documented to have a significant effect on a standardized output a key piece of evidence would have been brought forward to the educational equation. Formative assessment-driven instruction could then have been viewed as compatible

with the necessary large-scale quantifiable systems currently in place in the American public education system.

Definition of Terms

Formative assessment-driven instruction has articulated practical components for classroom use (Marzano, 2006; Moss & Brookhart; 2010; Popham, 2006; Tomlinson, 2014). The common elements of the formative assessment-driven instructional approach used by the instructor in this study synthesized the larger body of work of the approach to the following elements:

- course content that is guided by specific learning targets within traditional chapters/units,
- lessons organized in a backward fashion based on learning targets,
- grading structures divided between summative and formative work with summative work heavily weighted,
- formative assessment that includes traditional assignments, quizzes and practice tests as well as "in the moment" teaching adjustments,
- summative work that includes students doing a test corrective process where they self-assess their learning, determining errors where learning was not completed versus what was learned but incorrectly applied to a particular test question.

Assumptions and Limitations

This study was dependent on the fidelity of a single high school instructor using a formative assessment-driven instructional approach over the five years of this study. The students in the experimental group were consistently guided in their coursework by learning targets, formatively guided through their coursework in a backward design fashion, and

summatively assessed with a built-in revision system. It was assumed the instructor followed this methodology consistently.

It was also assumed that students gave an effort of the Minnesota Comprehensive Assessment for Science generally consistent with their student achievement level. The Minnesota Comprehensive Assessment for Science was not given the importance of Minnesota Comprehensive Assessments for Math and Reading. Those assessments were recorded on high school transcripts and could be used to assess college course assignments.

Limitations affecting this study included the use of a single instructor to define the experimental group. It was possible that the instructors for the control group courses adopted some formative processes through natural collegial collaboration. The instructors had rooms on the same floor and actively worked together in department meetings and collaborative teams. There were also changes in instructors in the control group as three different teachers left the school being studied. The teachers in turn were replaced by new hires.

Nature of the Study

This study used data retrieved from an electronic records system: Minnesota Comprehensive Assessment for Science scores and grade point averages. Students were categorized according to their science instructor.

Organization of the Remainder of the Study

The remainder of this study was divided into four chapters. Chapter Two reviews the literature on formative assessment as well as standardized testing. The case for formative assessment was detailed. The dual and competing narratives of standardized testing were then examined in regards to necessity and effectiveness. Chapter Three details the methodology that was used to determine if a statistical case can be made linking a formative assessment-driven instructional style with improved standardized test results. Chapter Four follows with the

findings from the study. Chapter Five includes conclusions, discussion and future considerations.

Chapter II: Literature Review

Introduction

This literature review examined the practice of formative assessment within an instructional approach. A formative assessment approach, when implemented according to researched based practices, engages students in an authentic learning process. Formative assessment draws on "best practice" instructional approaches with specific attention paid to the potential of feedback. The use of formative assessment-driven instruction has implications for grading practices. This chapter ends with a review of the practice of standardized testing detailing its advantages, complexities, criticisms, and compatibility with a formative assessment approach.

Formative Assessment Definition

Formative assessment is ongoing communication between teacher and student for the purpose of promoting learning. It typically contrasts with summative assessment. Shute (2008) described it as information communicated to the learner intended to modify their thinking with the purpose of improving learning. Marzano (2010) added that formative assessment communication is to be used by teachers to check the learning process for the purpose of informing decisions about future instruction. Popham (2006) further clarified:

> Formative assessment is simply a planned process wherein teachers, or their students, used assessment elicited evidence of student learning to decide whether to make changes in what they're currently doing. Formative assessment is assessment for learning as opposed to assessment of learning (p.4).

A formative assessment-driven instructional approach is dynamic in that student learning shapes instruction. "The primary purpose of formative assessment is to improve learning, not merely

audit it" (Moss & Brookhart, 2010, p. 58). Instructional practices are formative in classrooms when evidence about student achievement is elicited, interpreted, and used by teachers, learners, or peers to make next steps in instruction (Black & Wiliam, 1998a).

Formative Assessment Components

Improved student achievement is the goal of a formative assessment-driven instructional approach. Formative assessment needs to be continuous to produce a positive effect on student achievement. Bailey and Jakicic (2012) affirmed that teachers should regularly diagnose and assess learning for mastery within the classroom. Tomlinson (2014) suggested that "an ongoing exchange between a teacher and his or her students is designed to help students grow as vigorously as possible and to help teachers contribute to that growth as fully as possible" (p. 14). Black and Wiliam (1998b) cited extensive research where the use of formative assessment consistently produced an effect size between .04 and .07. They noted these effect sizes were larger than those of most educational interventions. They concluded that formative assessment practices had a positive effect on student achievement compared to systems based solely on summative assessments.

In terms of substance, Tomlinson (2014) itemized ten principles a formative assessment instructional approach should be based on: student understanding of the role of formative assessment, clear learning targets, accounting for student differences, instructive feedback, user friendly feedback, persistent use of formative assessment, student engagement with formative assessments, noticing patterns, planning instruction around content requirements and student needs, and repetitive use of formative assessment (p. 11-14). Within this structure, it is critical that students understand two elements in the formative assessment instructional process: what it is that they are to learn and how assessment will be used to achieve that learning. Those two

elements provide a foundation where a formative instructional approach will have success. Clearly communicated goal setting drives learning and achievement (Locke & Latham, 2002).

Formative assessment is fundamentally feedback based. Hattie and Timperley (2007) promoted feedback as an entity that enhances classroom learning as a whole. They stated that the main purpose of feedback is to reduce discrepancies between current understandings of performance and a goal (p. 86). Clark (2011) further clarified this: "formative feedback closes the gap between students' current level of understanding and the desired learning goal. It helps students understand the relationship between a clearly defined set of criteria or standards and their current level of performance" (p. 159).

Black and Wiliam (1998a) referenced the "purpose and placement" of content delivery within a formative assessment instructional model as a key ingredient for success. Considering purpose and placement make formative assessment useful for both the student and the teacher. Students receive feedback that helps them achieve their learning objectives. Teachers simultaneously gain insight as to how to instruct to meet student needs. Wiliam (2007) noted that when done effectively, formative assessment has the power to double the speed of student learning. Bailey and Jakcic (2012) contended that frequent and specific feedback deepens conversation around student learning. Students are able to make specific comparisons between their work and indicators of quality (p. 87-88).

Feedback loops originate from student to teacher and should be considered within a learning context. Feedback has no value when it exists in a vacuum (Hattie & Timperley, 2007 p. 82). Feedback needs to be timely. Feedback received after summative assessments comes too late in the learning process for it to be of value to students (Huxman, 2007). Hattie and Timperley (2007) provided a practical structure to acquire feedback:

> Effective feedback must answer three major questions asked by a teacher and/or student: Where am I going? (What are the goals?) How am I going? (What progress is being made toward the goal?) and Where to next? (What activities need to be undertaken to make better progress?) (p. 86).

Koenka and Anderman (2019) found that student-centered information delivered to students improved their performance. Feedback was most helpful when specific, task focused, not norm referenced, and not linked to personal characteristics (p. 15-22).

The feedback generated from a formative system improves the use of questioning. Wiliam (2014) assessed the drawback with the traditional routine questioning model. "Many students decline invitations to participate, random selection-oriented participation involves relatively few students, and teachers rarely plan their questioning" (p. 17). Druckor (2014) provided emphasis for the importance of questioning, calling for the development of all-student response systems (p. 18). Wiliam (2014) showed that it is particularly effective to forego questions entirely and instead make statements to which students are expected to respond (p. 18). This framework gives the practice of formative assessment legitimacy with students as it has the potential to connect with their lives outside of school.

Formative Assessment: Student - Teacher Interactions

The use of a formative assessment instructional approach blurs the traditional lines between instructional delivery and assessment. Assessment begins to drive instruction via a meaningful feedback flow. Feedback, when used for a correctional purpose, merges into the instructional process so thoroughly so that "the process itself takes on the forms of new instruction, rather than informing the student solely about correctness" (Kulhavy, 1977, p. 212). Black and William (1998a) concluded that instructional practices are formative in classrooms

when evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers to make next steps in instruction (p. 2).

Central to the success of a formative assessment-driven instructional approach is that the student simply has a better experience in it. Student motivation and effort increase when formative assessment is used to bridge learning gaps (Shute 2008). This system produces a classroom shift to a focus on learning instead of an anxious focus on grading (Wiliam, 2007). Students realize their own potential and strengths. Formative assessment emerged as a formative evaluation theory that focused on building off of student strengths. The underlying assumption is that virtually all human beings have dynamic potential (Scriven, 1967, p. 16).

When students receive feedback, teachers simultaneously gain insight as to how to instruct to meet student needs (Wiliam, 2007). A partnership mentality emerges producing trusted relationships. Druckor (2014) stated that formative assessment makes a difference not only for student outcomes but also for principals and teachers looking to build stronger relationships in their schools and classrooms. Clark (2011) stated that the interaction between students and teachers invariably involving peer collaboration enhances the educational process. Black and Wiliam (1998a) itemized the partnership process:

> Ultimately, emphasis on teacher-student interactions bring focused attention to the partnership aspect of learning. Classroom environment is required for these forces to thrive. Classrooms where implementation of formative assessment practice occur with fidelity are characterized by continuous assessment for learning, shared decision-making processes, clear learning targets, and both student and teacher monitoring of learning outcomes (p.7).

Formative Assessment: Theoretical Foundations

The formative approach creates an optimal learning environment. Vygotsky (1978) described this environment as a student's Zone of Proximal Development (ZPD). The collaboration and interaction of a formative approach mirrors Vygotsky's (1978) belief that an individual's full cognitive development requires social interaction as opposed to working in isolation. "Learning awakens a variety of internal developmental processes that are able to operate only when the child is interacting with people in his environment and in cooperation with peers" (p. 86).

Vygotsky's work contrasts with Piaget's constructivist view where discovery learning was seen as the basis for cognitive development. Piaget (1970) had found development to precede learning. Vygotsky (1978) felt social interaction and learning create conditions for cognitive development. Consciousness and cognition are the end product of socialization and social behavior.

Gutek (2011) blended the views of Vygotsky and Piaget to include cultural factors. "Focusing on the child alone tends to encourage us to look for causes of behavior with the child rather than the culture" (p. 171). Culture refers to a system of shared beliefs, values, knowledge, skills, relationships, customs, and practices (Gutek, 2011, p. 172). Natural socialization processes, in essence, provided a context for learning that could not be easily separated from social interaction or cognitive development. A formative approach, with its interactive ongoing use of diagnostic assessments, provides opportunity for inclusive learning environments for all students.

Yeager and Dweck (2012) described this process in the context of emerging brain research, with students embracing a growth mindset.

> We have found that what students need the most is not self-esteem boosting or trait labeling; instead, they need mindsets that represent challenges as things that they can take on and overcome over time with effort, new strategies, learning, help from others, and patience. When we emphasize people's potential to change, we prepare our students to face life's challenges resiliently. (p. 312).

Formative assessment decentralizes the learning process and provides students with more ownership of their learning.

Formative Assessment: Grading Considerations

Formative assessment-driven instructional approaches have the same finishing point as traditional instructional approaches (Sadler, 1998). Both are employed at the high school level towards end of course grades and credits for graduation. Summative assessments and grading structures are needed organizational entities for those systems to work. The use of a formative assessment-driven instructional approach does inject a new dynamic into the current high school model. Formative work needs to be accounted for in some fashion and blended with summative work for final course grades. The development of a formative assessment-driven instructional approach mandates the employment of a workable grading system. End of course results should not compromise the formative process. Chappuis (2014) added that an itemized process stemming from student feedback should not result in a low grade assigned too soon. Formative assessment-driven instructional approaches optimally provide students with ownership of their learning.

Bloom's (1968) promotion of mastery learning theory provides a helpful model for the formative process that moved learning theory into practice. Mastery learning is a clearly described level of top performance that becomes the standard of mastery for all students. With

sufficient time and skillful corrective instruction, Bloom believed that 95 percent of students could achieve mastery. Formative assessments were to be used along the way with feedback given as to whether mastery had been achieved. Students who had not achieved mastery were to receive diagnostic and prescriptive instruction from the teacher and additional chances to demonstrate mastery. In short, Bloom believed in comments to guide under-par performance to mastery grades, guided by clear expectations up front. Bloom's system necessitates the use of social interaction to guide learning and provides a blueprint for formative assessment as an instructional approach.

The higher student achievement promoted in a formative system has the potential to build student capacity. Sadler (1989) stated that "the instructional system must make explicit provision for students themselves to acquire evaluative expertise" (p. 143). Feedback is a consequence of performance. The quality, nature, and content of teachers' comments make a difference (Hattie & Timperley, 2007). Guskey (2019) suggested that feedback is central to any meaningful grades. In the end, "they are simply labels attached to different levels of student performance that describe in an abbreviated fashion how well students performed" (p. 45). He went on to add that the nature of the comments is the key factor. "Knowing where you are is essential to understanding where you need to go in order to improve" (p. 45). This metacognitive awareness also makes students better judges of their own work and increasingly self-sufficient as learners.

In further emphasizing that a formative assessment process requires a different grading mentality, Guskey (2019) noted that grades that compare students to their peers do not move learning forward. In fact, said Guskey, "Such competition is detrimental to relationships between students and has profound negative effects on the motivation of low-ranked students" (p. 46). Bloom (1968) had earlier mapped out grading guidelines that support the process promoted by

formative assessment noting areas of accomplishment, identification of improvement areas, and guidance on steps needed to effectively meet the learning criteria.

Standardized Testing Overview

Standardized tests have been employed in some fashion in United States' schools since the 1800's. They have been heavily employed to gather educational data in the school reform era of the past four decades. Standardized test scores became the core educational reporting tool after the passage of the 2002 No Child Left Behind Act. They soon became the most commonly used external measure of schools. Classroom assessments are given more broadly and frequently than standardized tests. These assessments are the most common measure used inside schools to measure student achievement and end of course grades.

Standardized tests provide common footing on what data should be collected in schools (Schneider, Feldman, & French, 2016). Ravitch (2010) stated that they can inform educational leaders and policy makers about the progress of the education system as a whole. Promoters of standardized tests find them to be reliable and objective measures of student achievement. Without them, policy makers would have to rely on tests scored by individual schools and teachers who have a vested interest in producing favorable results. Teacher subjectivity becomes a nonfactor. Standardized tests promote a sense of fairness in that they are inclusive and nondiscriminatory. School level results from standardized test scores reveal achievement disparities across race, gender, and income, protecting the interests of historically marginalized groups. Standardized tests represent meaningful student achievement and serve as a safeguard to social promotion (Phelps, 2002).

Standardized Testing Complexities

Other research presented a more complex picture where standardized testing results reinforce the status quo. A 2016 comprehensive study (White, et al.) suggested that marginalized groups will not be able to meaningfully experience equity in the current standardized testing system. "Thus, although reforms work to document progress with standardized test scores, these tests may be, in fact, measures of less mutable factors, such as race and SES, factors which may exert a compounding impact on achievement" (p. 10). Standardized test scores in their view, not only tended to reflect students' SES levels, they reinforced their impact at a school level. SES is one of the "strongest correlates of academic performance, although correlation at the school level were even stronger" (p. 11). Although SES has many operationalizations, it seems clear that high SES affords children an array of tangible and intangible supports that provide a developmental and lifelong benefit (Bradley & Corwyn, 2002). Some of the specific reform measures employed to improve standardized tests scores, such as school and class size, do produce significant effect sizes. But their collective gain is not enough to close the achievement gap.

Standardized testing is technologically dependent. The common use of multiple-choice questions on standardized tests that are graded by machine make them not subject to human subjectivity or bias. Technology does have inherent limitations in that it cannot adequately measure multiple types of student learning. But it is the hope for promoting not only accountability and instruction, but also a system that captures useful information while strengthening learning (Phelps, 2011).

Promoters state that frequent standardized tests have resulted in higher student achievement (Hauushek, 2014; Phelps, 2011). The tests ensure that basic skills are emphasized

in classrooms, eliminating time wasting activities. Standardized tests are not narrowing the curriculum, rather they are focusing it on important basic skills all students need to master. A 2005 study reported standardized testing had a positive impact, improving the quality of the curriculum while raising student achievement (Yeh, 2005). Conversely, others argue that standardized tests are an unreliable measure of student performance. The Brookings Institution reports (2012) found fifty % to eighty % of year-over-year test score improvements were temporary and caused by fluctuations that had nothing to do with long-term changes in learning (Whitehurst, 2014).

Standardized Testing Criticism

Critics of standardized testing cited numerous flaws with their use. They produce an emphasis on rote learning, encourage the elimination of curriculum deemed not central to test performance, pressurize the work and careers of teachers, and promote unnecessary competition. The system is a detriment to several types of learning styles. There is little room for creativity and imagination in a standardized focused world. Critical thinking is shorted in the standardized process. Critics say the standardized tests system has become a lucrative cottage industry attached to education. Standardized tests mostly benefit companies making millions from them (Koretz, 2017).

A common criticism leveled against standardized testing states that teachers are forced to teach to the test. Most teachers acknowledge the importance of standardized tests and do not feel their teaching has been compromised, according to a 2010 Gates Foundation study. A large majority (81%) of United States public school teachers said state-required standardized tests were at least "somewhat important" as a measure of students' academic achievement, and 27% said they were "very important " or "absolutely essential." Yeh (2005) found that teachers and

principals were widely aware that "isolated drills on the types of items expected on the test" were unacceptable (Yeh). Barth and Mitchell (2006) reported teaching to the test efforts to be unproductive.

> In any case, research has shown that drilling students does not produce test score gains: teaching a curriculum aligned to state standards and using test data as feedback produces higher test scores than an instructional emphasis on memorization and test-taking skills (p. 1-2).

Frey and Schmidt (2010) concluded that there were serious doubts as to whether classroom assessments produced more valued outcomes as opposed to standardized tests. They found the bulk of classroom assessments to be at Bloom's Taxonomy Levels One or Two with little critical thinking required.

Standardized Testing Proponents

Phelps (2011) reported that 93% of studies on student testing, including the use of large-scale and high-stakes standardized tests, found a "positive effect" on student achievement. A more complex result presented itself in 2016 poll data. A majority of public-school parents (58%) were confident that standardized tests did a good job of measuring how well their child was learning, but a mere 19% were very confident of this. Additionally, nearly half (49%) said standardized tests did not measure developmental life skills that were important to them. Less than half (39%) were confident standardized tests could measure those skills. An overwhelming majority (84%) said schools should assess these skills (Phi Delta Kappan, 2017). Transfer abilities are increasingly in demand in the workforce. Transfer abilities can be best measured through authentic, performance-based tasks, with well-developed rubrics for evaluation **(**McTighe**,** 2018).

Most students believe standardized tests are fair. A 2006 survey of public-school students in Grades 6-12 found that 71% of students think the number of tests they have to take is "about right" and 79% believe test questions are fair. An earlier version of the study (2002) found that "virtually all students say they take the tests seriously and more than half (56 %) say they take them very seriously (Wang, Gulbahar, & Brown, 2006, p. 305-306).

Standardized tests hamper multiple types of student learners such as those with testing anxiety, or those needing extended reflective time to respond to complex scenarios. Proponents of standardized testing believe the testing anxiety issue to be limited and within margins of acceptability. The U.S. Department of Education (2014) stated: "Although testing may be stressful for some students, testing is a normal and expected way of assessing what students have learned" (p.1). The study found that "the vast majority of students do not exhibit stress and have positive attitudes towards standardized testing programs" (p.1.).

Standardized Testing Results

Throughout the ongoing debate on the merits of standardized testing, there is consensus that the results have not been adequate given the retooled goals of United States education. The two major international comparison entities are the National Assessment of Educational Progress (NAEP) and Trends in International Mathematics and Science Study (TIMSS). The congressionally mandated NAEP, known as the “nation’s report card,” has provided information about student performance since 1969. It is the only assessment that measures what U.S. students know and can do in various subjects across the nation, states, and in some urban districts. Scores are compiled for multiple school subjects. Mathematics and reading scores are used as a common denominator to gauge academic achievement in U.S. schools. TIMSS provides data on the mathematics and science achievement of U.S. students compared to that of

students in other countries. TIMSS data have been collected from students at Grades 4 and 8 since 1995 every four years, generally. In addition, TIMSS Advanced measures advanced mathematics and physics achievement in the final year of secondary school across countries. TIMSS Advanced data has been collected internationally three times, in 1995, 2008 and 2015.

The NAEP scores tabulated in the 1990's painted a bleak picture of achievement in U.S. schools. NAEP has set a standard for American students that the majority of students in the world cannot meet (Loomis & Bourque, 2001). NAEP data over subsequent decades does show longitudinal improvement in reading and math since 1990. Scores declined modestly in all reading and math areas from 2017 to 2019 with the exception of Grade four mathematics (U.S. Department of Education, 2019). TIMMS data shows U.S. students showing little or no growth since 1995. Several Asian nations have surpassed the United States in overall TIMMS achievement in that time span (U.S. Department of Education, 2019).

While standardized test scores have limitations as a measurement of student learning, they are a necessary component of any accountability system done on a significant scale. Standardized tests are time bound. The reliability and validity expectations necessitate appropriate security and implementation systems that yield a pressurized single setting testing environment. Over 30 million students attend school in the United States, sprawling across fifty states plus the District of Columbia (U.S. Department of Education, 2019). A mechanism to produce comparable, valid data on such a scale makes the standardized test central to the determination of educational results. When considering the massive task of measuring the holistic achievement of millions of students, standardized tests are the only option to produce a concise summary.

Summary

For the United States public school system to produce an overall level of student achievement that is acceptable to the various stakeholders dependent on the educational system, two criteria must be addressed. Quantifiable results must be produced that reflect the vast and complex student needs of United States public school students. Effective classroom instruction must occur on a widespread basis in a manner that reflects researched based practices to produce those high student achievement yields.

Chapter III: Methodology

Introduction

This study leveraged an opportunity to investigate the depth of interaction between formative assessment-driven instruction and standardized test scores. Minnesota introduced a standardized test in 2008, the Minnesota Comprehensive Assessment for Science (MCA). The MCA was based on the Minnesota Academic Standards for Science addressed in ninth and tenth grade science courses. The test helped districts measure student progress towards proficiency or mastery of standards. The Minnesota Comprehensive Assessment for Science is unique from other Minnesota standardized tests in that it is based on academic standards tied to specific courses at the high school level. The standards have a cursory introduction in ninth grade and then receive full scale emphasis in tenth grade.

One science instructor did a full-scale revision of this teaching methodology to a formative assessment-driven instructional approach in 2008. The instructor's students took a tenth grade biology course and then took the same Minnesota Comprehensive Assessment for Science as students taking biology from other instructors via a traditional instructional approach. That instructor's students became the experimental group in this study with the students receiving traditional inspection serving as the control group. The groups were used to assess the depth of interaction between formative assessment-driven instruction and standardized test scores.

Purpose of the Study

The efforts to reform the education system of the United States have received prioritized funding since the world moved into the global era. Educational funding comes from a mix of federal, state, and local sources. Measured as a percentage of the Gross Domestic Product

(GDP) of the United States, it grew roughly a full percentage point from the mid 1980's to the mid 2010's. Educational spending was then impacted by the 2008 economic downturn and has since then struggled to return to pre-2008 levels. The educational funding equation is complex with federal mandates placing requirements on state and local expenditures. The requirements have included a host of standardized testing and data collection components (Leachman, Masterson, & Figuero, 2017). Technological advancements have produced new capabilities to collect and analyze educational data for the purpose of improving instruction. The wide array of data collected has helped to develop and validate improved instructional practice as well as serving accountability purposes on large scale state and federal levels.

The large-scale data usage has been nearly universal in its form: standardized test results. The accountability system measures student learning based on articulated academic standards. The results have been mixed at best. A consensus within the United States is that multiple waves of reform, highlighted by the No Child Left Behind Act of 2002 and its 2015 Every Student Succeeds Act successor, have not adequately produced an overall achievement level in United States public schools deemed to be internationally competitive. Embedded in the push to raise student achievement was the elimination of the "opportunity gap" that exists between different races or ethnicities within the United States. Data has shown time and again that white students consistently outperform non-white students across the nation when it comes to standardized test results.

Data generated at the classroom level showed great promise for the practice of formative assessment-driven instruction. The purpose of this study was to investigate how that method of instructional practice correlated to standardized test scores. The question was what is the

relationship between formative assessment-driven instruction and improved student performance on standardized tests, particularly for average or below average students?

Conceptual Framework

Accountability measures have relied on educational output data, namely standardized test scores, to gauge educational success. Input measures of the educational success equation have also received focus during this time span. A great deal of resources has been directed or redirected towards the output desired goals of improved student achievement and the elimination of the opportunity gap. Instructional practice lies between the input and output portions of the education equation. The attention paid to the instructional process has been minimal in comparison to input and output measures. Improved pedagogical processes have been developed during the reform efforts in the shadows of larger and more public narratives. The areas of brain research, instructional approaches, and learning styles have all contributed to a more enlightened and effective instructional model. They have collectively combined to produce improved direction to spur better student achievement. Research has shown (Bailey & Jakcic, 2012; Black & Wiliam, 1998; Frey & Schmidt, 2010) that the use of assessment to drive instruction, commonly labeled formative assessment, has consistently produced improved student achievement.

Research Design

This was a quantitative ex-post facto archival case study, utilizing a non-equivalent control group design. This research effort examined five years of data to determine whether high school students who had been exposed to formative assessment teaching processes correlated to better standardized test results than peers who have not had that exposure. Two groups were compared on their Minnesota Comprehensive Assessment (MCA) for Science performance. One

of the groups had been exposed to a formative assessment-driven instructional model. This group was the experimental group subject to the independent variable. The other non-exposed group was the baseline, control group. The purpose of the research was the determination of the relationship between formative assessment-driven instruction and student performance on the Minnesota Comprehensive Assessment for Science.

An investigation of the relationship of formative assessment-driven teaching and MCA for Science scores was determined first by comparing the performance of all students on the MCA for Science to their overall level of student achievement. Students were placed in four quartiles according to their overall high school grade point averages. Since better students commonly score better on standardized tests, it was speculated that some students score well on the MCA for Science without regard to a specific approach to instruction. Student performance on the MCA for Science was then reviewed in the control and experimental groups given their respective student achievement quartiles.

Research Question

The research question for this study was: what is the relationship between formative assessment-driven instruction and standardized test scores, particularly for average or below average students?

Hypotheses

Hypothesis One: There will be significant differences in performance on the Minnesota Comprehensive Assessment for Science scores between students receiving formative assessment-driven instruction and students receiving traditional instruction.

Null Hypothesis One: There will be no differences in the Minnesota Comprehensive Assessment for Science scores between students receiving formative assessment-driven instruction and students receiving traditional instruction.

Hypothesis Two: Student Minnesota Comprehensive Assessment for Science scores will correlate with their overall student achievement level as measured by quartiles of four-year grade point averages.

Null Hypothesis Two: Student Minnesota Comprehensive Assessment for Science scores will not correlate with their overall student achievement level as measured by quartiles of four-year grade point averages.

Hypothesis Three: There will be a significant interaction between students receiving formative assessment-driven instruction and student achievement quartiles.

Null Hypothesis Three: There will be no significant interaction between students receiving formative assessment-driven instruction and student achievement quartiles.

Variables

The two independent variables in this study were the type of instruction (formative assessment-driven vs. traditional) and student academic achievement (GPA). Student academic achievement was measured by putting all students into quartiles based on their GPAs. The dependent variable in this study was Minnesota Comprehensive Assessment (MCA) Science scores.

Instruments and Measures

This was a quantitative ex-post facto archival case study. Student data used were the scores from the Minnesota Comprehensive Assessment for Science over a five-year period from 2009 to 2013. Students took the Minnesota Comprehensive Assessment for Science during the

spring of their sophomore years. Four-year grade point averages for the same students graduating from 2011 to 2015 were the second piece of data used.

Sampling Design

The subjects in this study attended a Minnesota inner ring suburban high school from a major metropolitan area. Approximately 1,950 students attended the high school. The majority of the students resided in one of three communities. Approximately ten % of the students from the high school were open enrolled from other school districts from the metropolitan area. The high school saw a steady diversification of its student body during the time of this study. Students of color accounted for 36% of its student body in 2008-2009. By year five of the study students of color were 48% of the student population. Free and reduced lunch students mirrored the growth of students of color growing from 40% in 2008-2009 to 48% in 2012-2013. The students of color who attend the high school were primarily of Black and Asian races. The number of Hispanic/Latinx students increased during the five-year time period of the study.

All subjects used in the study took a full year biology course at said high school. Biology was a required course for sophomores and is required for graduation. Only full year students were included in this study. Transfer students completing less than the full year of study were excluded. Students with unique special education or English learner needs were also excluded from this study. These exclusions were made to reduce potential reliability and validity barriers.

The subjects in this study were high school sophomores who enrolled in a required biology course over a five-year period from 2009 to 2013. The graduation year for these students was 2011 to 2015. This study was limited to five years as the distinctions between the control and experimental group began to blur as the other science teachers as well as most other teachers in the school adopted formative assessment-driven formative processes. There was no

longer a unique group of students during the 2013-2014 school year being instructed in a formative assessment-driven manner.

A different version of the Minnesota Comprehensive Assessment for Science was given during the last two years of this study. The Minnesota Comprehensive Assessment for Science II was given from 2008 to 2011. The Minnesota Comprehensive Assessment for Science III was given beginning in 2012. An adjustment in statewide scores occurred as a result of the new test. All subjects included in this study took the same version of the Minnesota Comprehensive Assessment for Science in each given year of the study.

Group selection was subject to the parameters of the master scheduling calendar of said high school. The science courses taught in ninth and tenth grade were the academic foundations upon which the Minnesota Comprehensive Assessment for Science was based. Ninth graders were assigned three trimesters of earth science and tenth graders three trimesters of biology. Students were placed in common courses. There were no advanced or accelerated options for students. Teachers were assigned to teach sections of science courses per license and scheduling needs. Students were placed in sections randomly.

The biology course being used in this study was taught over three trimesters. Biology teachers were classified as formative or traditional teachers. Students most often had the same instructor through all three trimesters of the school year. However, some students had a mix of teachers due to schedule parameters. For the purposes of this study, those students who were instructed in two or three trimesters in a formative manner were included in the experimental group. Those formatively instructed in zero or one trimesters were included in the control group.

The students in this study took the Minnesota Comprehensive Assessment for Science during the last month of the same school year they took the full year biology course. The

number of students involved each year ranged from 325 to 413. Multiple individuals instructed the biology courses during these years. The number of sections taught each year varied from 13 to 16 depending on enrollment. The instructors involved in teaching the biology course were classified as formative assessment or traditional teachers.

The common elements of the formative assessment-driven instructional approach used by the instructor in this study synthesized the larger body of work of the approach to the following elements:

- course content guided by specific learning targets within traditional units,
- lessons organized in a backward fashion based on learning targets,
- grading structures divided between formative and summative work, with summative work heavily weighted in course grades,
- formative assessments include traditional assignments, quizzes, and practice tests as well as "in the moment" teaching adjustments,
- summative work that includes students doing a test correctives process where they self-assess their learning determining errors where learning was not completed or was incorrectly applied.

Traditional teachers were not subject to any consideration of their instructional approaches.

Data Collection Procedures

The data used in this study was anonymous. The subjects were high school sophomores taking a full year biology course. The subjects attended a Minnesota inner ring diverse suburban high school from a major metropolitan area. Data used in this study was collected by an independent third party. The data used was retrieved from an Infinite Campus student information system.

Data Analysis

A 2 (formative assessment instruction vs. traditional instruction) X 4 (high, above average, below average, and poor GPA) factorial ANOVA was used to analyze MCA-Science scores. For the academic achievement variable, all students were placed into one of four quartiles based on their final high school GPA. All data was analyzed using SPSS version 26. Results were analyzed both collectively and on a year-by-year basis.

Reliability, Validity, and Trustworthiness

This study can be considered to be internally reliable within its scope of study. The instructional assignments of the teachers remained mostly constant over the five years of the study. The five years of data with each year involved four, five, or six teachers each year. The students involved in the study were divided into 16, 16, 15, 14 and 13 sections respectively over the five-year time period. The experimental group consisted of five sections each year taught by the same instructor. The remaining control group sections were taught by three, four or five teachers each year. There were multiple changes in the group of teachers in the control group over the five-year time period.

All students involved in the study took a common standardized test, the Minnesota Comprehensive Assessment for Science. A new version of the standardized test was given beginning in year four of the study. The instructional format remained constant for both the experimental and control groups over the five years of the study.

The use of students' grade point averages supported construct validity. The dynamics of formative instruction were assessed against student performance over time. Therefore, the results of this study can be generalized for application to other school subjects. The study can also be considered to have content validity as it covers a five-year period of time.

Limitations and Assumptions

This study was limited to a single high school, making it highly dependent on a finite number of teachers. It is possible that a number of different variables not related to formative assessment practices could impact findings. Among these variables are teacher availability and capability. It would be preferable to have had a baseline of the interaction of grade point averages with Minnesota Comprehensive Assessment for Science over multiple student populations and years. Each student could have been assigned an "expected success" score against which to measure the correlation of a newly implemented formative assessment-driven process. Minnesota Comprehensive Assessments were first given in 2003 for reading and mathematics across multiple grade levels.

The students in this study were quite familiar with the notion of a Minnesota Comprehensive Assessment. Minnesota Comprehensive Assessments for Science were first given in 2008. The initial version was given for four years and then updated for 2012. Statewide scores saw an adjustment that year.

The formative assessment adaptation of teacher practice was underway during the 2007-2008 school year. While a formal adoption of the system by one teacher was at first not adopted by colleagues, it is possible that some of the practices were adopted due to collaborative practice.

It is also possible that some students had altered mentalities towards learning as teachers in other departments experimented with and implemented formative assessment-driven instruction. Those students have had a higher capacity due to this. The findings of this study may not be generalizable to a larger population due to the difficulties of isolating the independent variable.

The nuances of high school scheduling invariably impact the makeup of individual classes. With each student taking six courses each school day, concentrated electives such as band tend to cluster students together in multiple other courses. With multiple required courses having accelerated or honors level sections, many students who take band end up having very similar daily schedules. These similar schedules may land students more proportionately in either the control or experimental group. Other tracking tendencies could also have impacted student schedules in this study.

Ethical Considerations

The author of this study was the lead administrator of the institution the data was retrieved from. This person was a promoter of the adoption of a formative assessment-driven instructional process but did not mandate its use. A spirit of experimentation was present within the teaching staff of the high school of this study as many teachers considered the merits of its use. Basic formative assessment elements were instituted school wide in the fourth and fifth years of the study. Teachers were required to post learning targets and use the categories of formative and summative in the grading procedures.

Two individuals other than the author retrieved and analyzed the data used in this study. The data was retrieved from the school's learning management system. The author was not involved in the input of the data.

Research has consistently shown that formative assessment-driven instruction processes improve student achievement. Given the solid endorsement of educational research, it made sense to see if entering formative assessment-driven instruction into the accountability equation was helpful to the overall goal of improved standardized test scores. If the substance between massive resource inputs and standardized test scores outputs could have been articulated in a

fashion that improves student achievement, an encouraging element would have been added to the overall current educational equation.

Chapter IV: Results

Overview

SPSS version 26 was used for all statistical analyses. A 2 (Formative Assessment Driven vs. Traditional) X 4 (GPA Quartiles) factorial ANOVA was used to analyze Minnesota Comprehensive Assessment scores (II or III depending on the year). Five years (2011-2015) of data were analyzed, which means five factorial ANOVAs were conducted in all. Tukey's HSD was used as a post hoc test to analyze any mean differences for the GPA Quartile variable when the overall *F* value for that variable was significant.

2011 (MCA-II)

There was a significant main effect for type of instruction. Students in the formative assessment driven class ($M = 1053.16$, $SD = 9.76$) scored significantly higher than students in the traditional teaching class ($M = 1050.17$, $SD = 9.32$), $F\ (1,405) = 5.43$, $p = .02$, $\eta^2 = .013$. As expected there was also a significant main effect for GPA, $F\ (3,405) = 43.73$, $p < .001$, $\eta^2 = .245$ (See Table 2 for full ANOVA table.) Tukey post hoc tests revealed that each GPA quartile group scored significantly different from each other (see Table 1 for means and standard deviations and Table 3 for post hoc results). Figure 1 demonstrates that these mean differences were linear. The figure also demonstrates that there was no significant interaction between the two independent variables, $F\ (3,405) = 0.56$, $p = .64$.

Table 1

2011 Means and Standard Deviations for MCA-II Scores by Class Type and GPA Quartiles

Class	Percentile Group of Final GPA	Mean	Standard Deviation	N
Formative Assessment Driven	Lowest GPA Quartile	1044.47	4.824	15
	Second Lowest GPA Quartile	1049.11	8.543	19
	Second Highest GPA Quartile	1054.12	7.512	26
	Top GPA Quartile	1061.09	9.175	23
	Total	1053.16	9.764	83
Traditional	Lowest GPA Quartile	1042.58	9.729	78
	Second Lowest GPA Quartile	1048.43	7.384	80
	Second Highest GPA Quartile	1051.74	7.289	84
	Top GPA Quartile	1056.98	6.406	88
	Total	1050.17	9.324	330
Total	Lowest GPA Quartile	1042.88	9.125	93
	Second Lowest GPA Quartile	1048.56	7.578	99
	Second Highest GPA Quartile	1052.30	7.378	110
	Top GPA Quartile	1057.83	7.217	111
	Total	1050.77	9.478	413

Table 2

2 (Class) X 4 (GPA Quartiles) ANOVA Table for 2011 MCA-II Scores

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	12532.901[a]	7	1790.414	29.622	.000
Class	328.343	1	328.343	5.432	.020
GPA Quartiles	7929.128	3	2643.043	43.729	.000
Class * GPA Quartiles	101.926	3	33.975	.562	.640
Error	24478.784	405	60.441		
Corrected Total	37011.685	412			

a. R Squared = .339 (Adjusted R Squared = .327)

Table 3

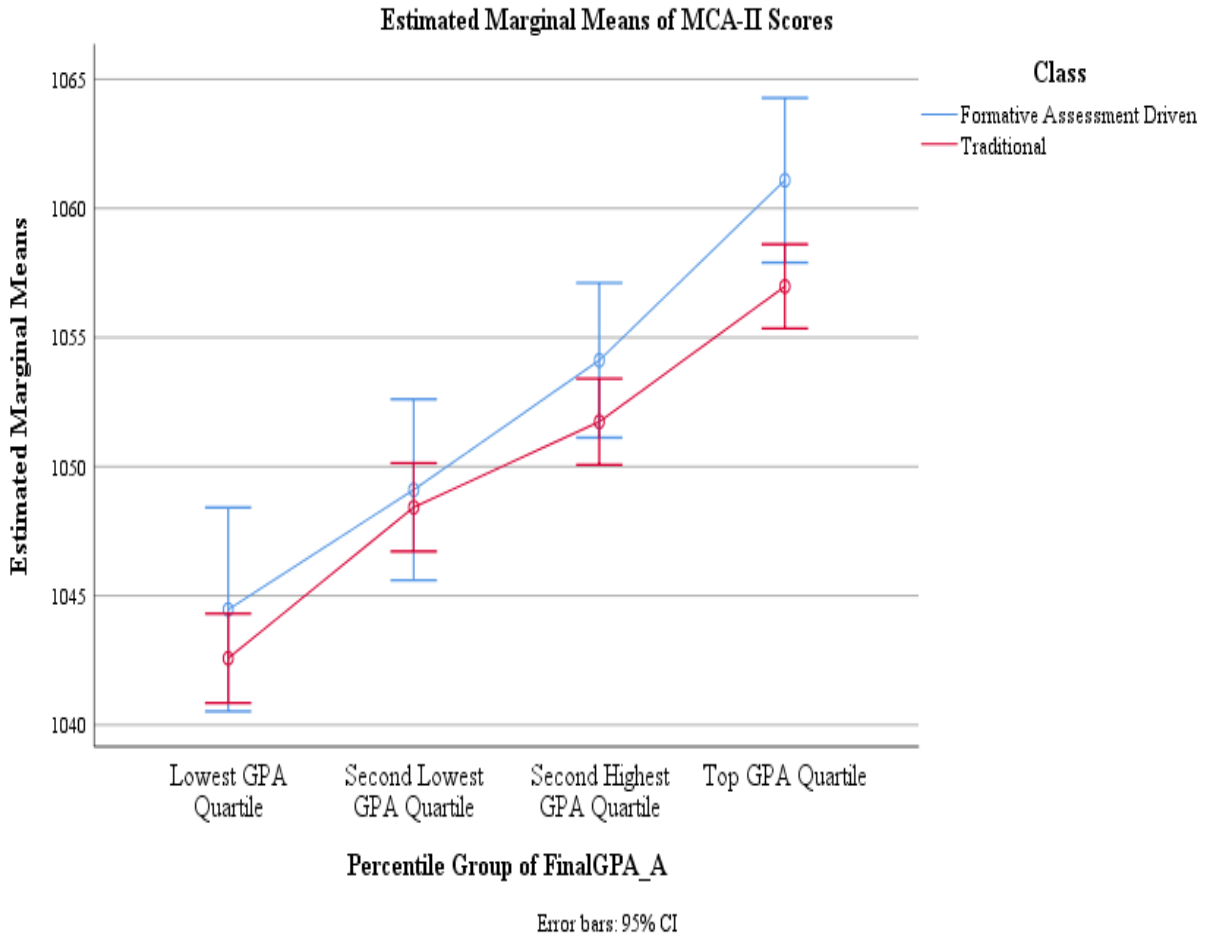
Tukey's HSD Post Hoc Tests for 2011 MCA-II Scores by GPA Quartiles

(I) Percentile Group of Final GPA	(J) Percentile Group of Final GPA	Mean Difference (I-J)	Standard Error	Sig.
Lowest GPA Quartile	Second Lowest GPA Quartile	-5.67*	1.123	.000
	Second Highest GPA Quartile	-9.42*	1.095	.000
	Top GPA Quartile	-14.95*	1.093	.000
Second Lowest GPA Quartile	Lowest GPA Quartile	5.67*	1.123	.000
	Second Highest GPA Quartile	-3.74*	1.077	.003
	Top GPA Quartile	-9.27*	1.075	.000
Second Highest GPA Quartile	Lowest GPA Quartile	9.42*	1.095	.000
	Second Lowest GPA Quartile	3.74*	1.077	.003
	Top GPA Quartile	-5.53*	1.046	.000
Top GPA Quartile	Lowest GPA Quartile	14.95*	1.093	.000
	Second Lowest GPA Quartile	9.27*	1.075	.000
	Second Highest GPA Quartile	5.53*	1.046	.000

*. The mean difference is significant at the .05 level.

Figure 1

2011 MCA-II Scores by Class Type and GPA Quartiles



2012 (MCA-II)

There was no significant main effect for type of instruction, F (1,400) = 0.29, p = .65. Students in the formative assessment driven class (M = 1050.28, SD = 8.98) performed similarly to the students in the traditional teaching class (M = 1049.59, SD = 9.36). There was a significant main effect for GPA, F (3,400) = 58.61, p < .001, η^2 = .305. (See Table 5 for full ANOVA table.) Tukey post hoc tests revealed that all groups were significantly different from each other, except for the second and third quartile groups (see Table 4 for means and standard deviations and Table 6 for post hoc results). Figure 2 demonstrates that these mean differences were linear. The figure also demonstrates that there was no significant interaction between the two independent variables, F (3,400) = 0.31, p = .82.

Table 4

2012 Means and Standard Deviations for MCA-II Scores by Class Type and GPA Quartiles

Class	Percentile Group of Final GPA	Mean	Standard Deviation	N
Formative Assessment Driven	Lowest GPA Quartile	1042.58	8.420	26
	Second Lowest GPA Quartile	1047.42	5.291	24
	Second Highest GPA Quartile	1050.43	5.399	30
	Top GPA Quartile	1058.81	7.314	31
	Total	1050.28	8.975	111
Traditional	Lowest GPA Quartile	1042.64	9.858	67
	Second Lowest GPA Quartile	1047.83	7.159	81
	Second Highest GPA Quartile	1050.10	7.277	73
	Top GPA Quartile	1057.12	7.105	76
	Total	1049.59	9.358	297
Total	Lowest GPA Quartile	1042.62	9.433	93
	Second Lowest GPA Quartile	1047.73	6.756	105
	Second Highest GPA Quartile	1050.19	6.759	103
	Top GPA Quartile	1057.61	7.173	107
	Total	1049.78	9.250	408

Table 5

2 (Class) X 4 (GPA Quartiles) ANOVA Table for 2012 MCA-II Scores

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	11844.516[a]	7	1692.074	29.456	.000
Class	12.023	1	12.023	.209	.648
GPA Quartiles	10101.140	3	3367.047	58.614	.000
Class * GPA Quartiles	52.686	3	17.562	.306	.821
Error	22977.631	400	57.444		
Corrected Total	34822.147	407			

a. R Squared = .340 (Adjusted R Squared = .329)

Table 6

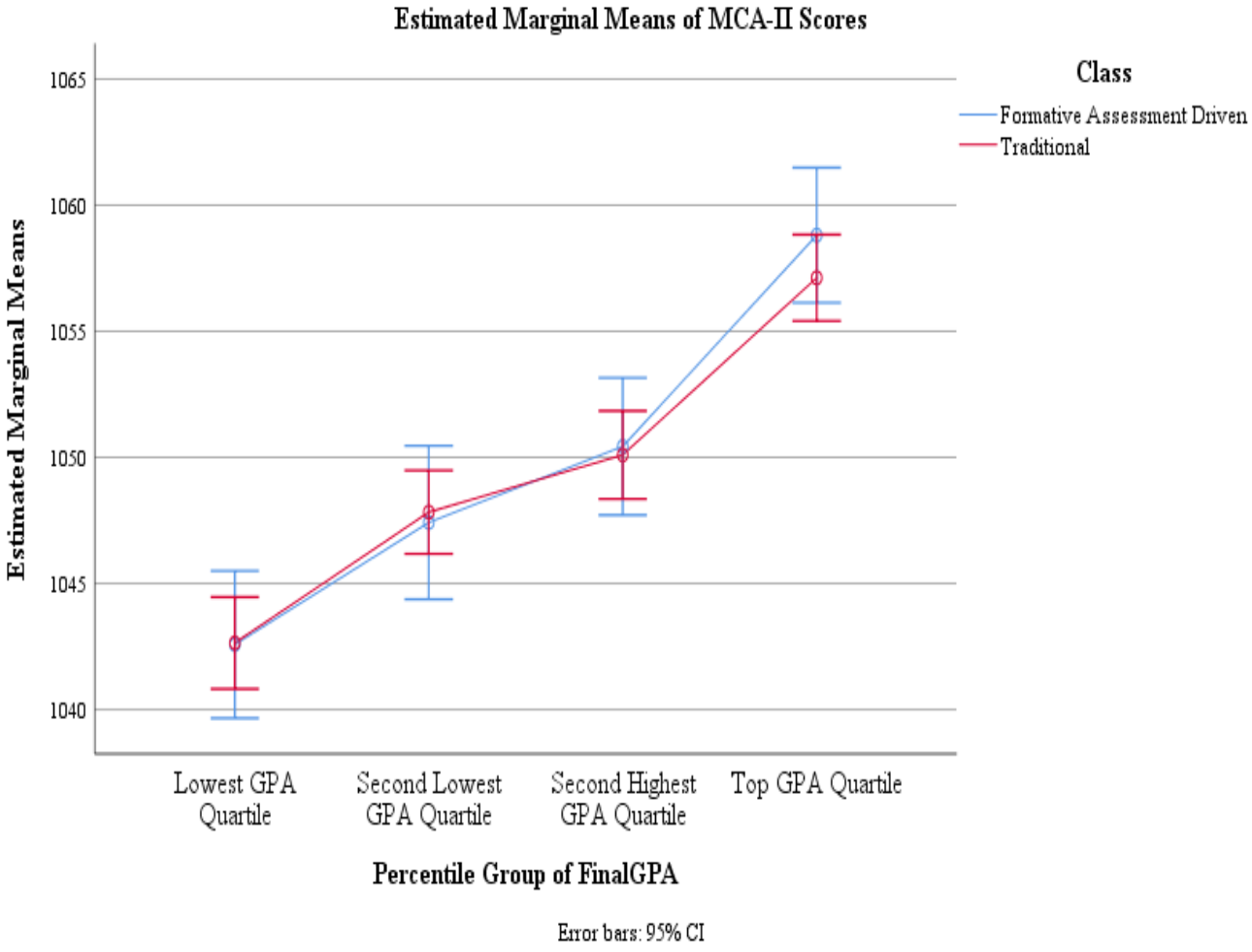
Tukey's HSD Post Hoc Tests for 2012 MCA-II Scores by GPA Quartiles

(I) Percentile Group of Final GPA	(J) Percentile Group of Final GPA	Mean Difference (I-J)	Standard. Error	Sig.
Lowest GPA Quartile	Second Lowest GPA Quartile	-5.11*	1.079	.000
	Second Highest GPA Quartile	-7.57*	1.084	.000
	Top GPA Quartile	-14.98*	1.074	.000
Second Lowest GPA Quartile	Lowest GPA Quartile	5.11*	1.079	.000
	Second Highest GPA Quartile	-2.46	1.051	.091
	Top GPA Quartile	-9.87*	1.041	.000
Second Highest GPA Quartile	Lowest GPA Quartile	7.57*	1.084	.000
	Second Lowest GPA Quartile	2.46	1.051	.091
	Top GPA Quartile	-7.41*	1.046	.000
Top GPA Quartile	Lowest GPA Quartile	14.98*	1.074	.000
	Second Lowest GPA Quartile	9.87*	1.041	.000
	Second Highest GPA Quartile	7.41*	1.046	.000

Based on observed means.
The error term is Mean Square (Error) = 57.444.
*. The mean difference is significant at the .05 level.

Figure 2

2012 MCA-II Scores by Class Type and GPA Quartiles



2013 (MCA-II)

Similar to the 2012 results, there was no significant main effect for type of instruction, F (1,387) = 2.96, p = 09. Students in the formative assessment driven class (M = 1047.55, SD = 9.83) performed similarly to the students in the traditional teaching class (M = 1049.63, SD = 10.06). There was a significant main effect for GPA, F (3,387) = 54.24, $p < .001$, $\eta^2 = .296$. (See Table 8 for full ANOVA table.) Tukey post hoc tests revealed that all groups were significantly different from each other (see Table 7 for means and standard deviations and Table 9 for post hoc results). Figure 3 demonstrates that these mean differences were linear. The figure also demonstrates that there was no significant interaction between the two independent variables, F(3,387) = 0.36, p = .78.

Table 7

2013 Means and Standard Deviations for MCA-II Scores by Class Type and GPA Quartiles

Class	Percentile Group of Final GPA	Mean	Standard Deviation	N
Formative Assessment Driven	Lowest GPA Quartile	1040.04	7.589	28
	Second Lowest GPA Quartile	1043.50	8.772	24
	Second Highest GPA Quartile	1050.93	8.467	27
	Top GPA Quartile	1055.56	6.047	27
	Total	1047.55	9.833	106
Traditional	Lowest GPA Quartile	1041.55	11.331	62
	Second Lowest GPA Quartile	1046.58	8.986	72
	Second Highest GPA Quartile	1051.24	6.193	74
	Top GPA Quartile	1057.05	6.548	81
	Total	1049.63	10.063	289
Total	Lowest GPA Quartile	1041.08	10.294	90
	Second Lowest GPA Quartile	1045.81	8.988	96
	Second Highest GPA Quartile	1051.16	6.830	101
	Top GPA Quartile	1056.68	6.431	108
	Total	1049.07	10.032	395

Table 8

2 (Class) X 4 (GPA Quartiles) ANOVA Table for 2013 MCA-II Scores

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	13718.253[a]	7	1959.750	29.245	.000
Class	197.999	1	197.999	2.955	.086
GPA Quartile	10904.707	3	3634.902	54.242	.000
Class * GPA	72.446	3	24.149	.360	.782
Error	25933.762	387	67.012		
Corrected Total	39652.015	394			

a. R Squared = .346 (Adjusted R Squared = .334)

Table 9

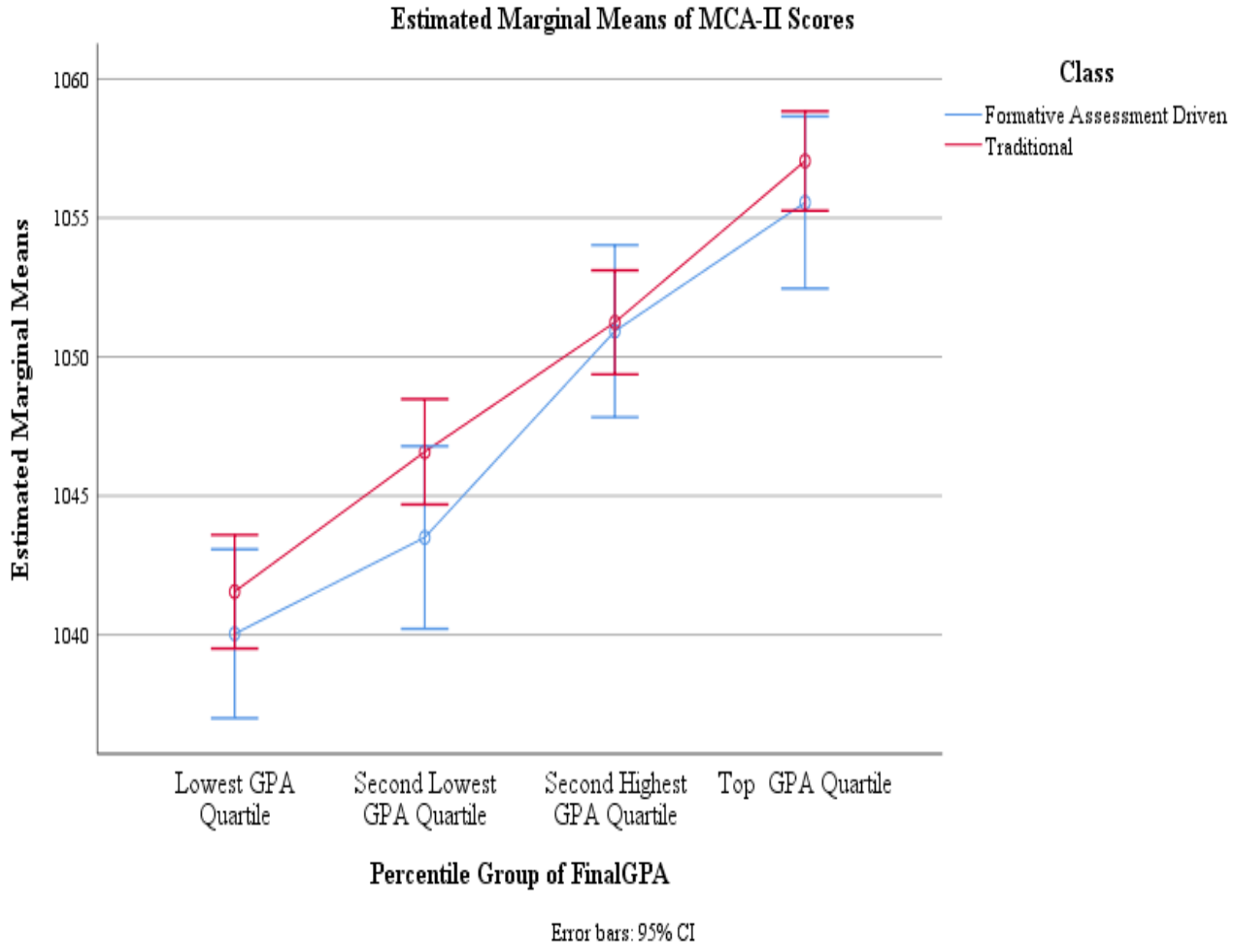
Tukey's HSD Post Hoc Tests for 2013 MCA-II Scores by GPA Quartiles

(I) Percentile Group of Final GPA	(J) Percentile Group of Final GPA	Mean Difference (I-J)	Standard Error	Sig.
Lowest GPA Quartile	Second Lowest GPA Quartile	-4.73*	1.201	.001
	Second Highest GPA Quartile	-10.08*	1.187	.000
	Top GPA Quartile	-15.60*	1.168	.000
Second Lowest GPA Quartile	Lowest GPA Quartile	4.73*	1.201	.001
	Second Highest GPA Quartile	-5.35*	1.167	.000
	Top GPA Quartile	-10.86*	1.148	.000
Second Highest GPA Quartile	Lowest GPA Quartile	10.08*	1.187	.000
	Second Lowest GPA Quartile	5.35*	1.167	.000
	Top GPA Quartile	-5.52*	1.133	.000
Top GPA Quartile	Lowest GPA Quartile	15.60*	1.168	.000
	Second Lowest GPA Quartile	10.86*	1.148	.000
	Second Highest GPA Quartile	5.52*	1.133	.000

*. The mean difference is significant at the .05 level.

Figure 3

2013 MCA-II Scores by Class Type and GPA Quartiles



2014 (MCA-III)

As was the case with the 2012 and 2013 results, there was no significant main effect for type of instruction, F (1,317) = 1.43, p = 23. Students in the formative assessment driven class (M = 1051.34, SD = 11.90) performed similarly to the students in the traditional teaching class (M = 1051.74, SD = 10.30). There was a significant main effect for GPA, F (3,317) = 46.03, p < .001, η^2 = .303. (See Table 11 for full ANOVA table.) Tukey post hoc tests revealed that all groups were significantly different from each other, except for the two lowest GPA groups (see Table 10 for means and standard deviations and Table 12 for post hoc results). However, in this data set there was a significant interaction, F (3,317) = 3.01, p = .03, η^2 = .028. An examination of Figure 4 reveals the different patterns of mean scores in the lowest three GPA groups by instructional type. In the lowest GPA group the traditional instruction group did best. In the second lowest GPA group the formative assessment group did best. But then again in the second highest GPA group the traditional group did better. In the highest GPA group there were similar scores between traditional and formative assessment groups.

Table 10

2014 Means and Standard Deviations for MCA-III Scores by Class Type and GPA Quartiles

Class	Percentile Group of Final GPA	Mean	Standard Deviation	N
Formative Assessment Driven	Lowest GPA Quartile	1040.86	14.708	21
	Second Lowest GPA Quartile	1049.87	9.108	30
	Second Highest GPA Quartile	1050.23	5.895	22
	Top GPA Quartile	1060.37	8.530	32
	Total	1051.34	11.901	105
Traditional	Lowest GPA Quartile	1045.59	9.506	51
	Second Lowest GPA Quartile	1046.66	9.348	56
	Second Highest GPA Quartile	1054.16	8.085	58
	Top GPA Quartile	1060.05	7.004	55
	Total	1051.74	10.296	220
Total	Lowest GPA Quartile	1044.21	11.370	72
	Second Lowest GPA Quartile	1047.78	9.339	86
	Second Highest GPA Quartile	1053.07	7.714	80
	Top GPA Quartile	1060.17	7.553	87
	Total	1051.61	10.823	325

Table 11

2 (Class) X 4 (GPA Quartiles) ANOVA Table for 2014 MCA-III Scores

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	12538.624[a]	7	1791.232	22.340	.000
Class	114.266	1	114.266	1.425	.233
GPA Quartiles	11071.032	3	3690.344	46.026	.000
Class * GPA	724.023	3	241.341	3.010	.030
Error	25416.748	317	80.179		
Corrected Total	37955.372	324			

a. R Squared = .330 (Adjusted R Squared = .316)

Table 12

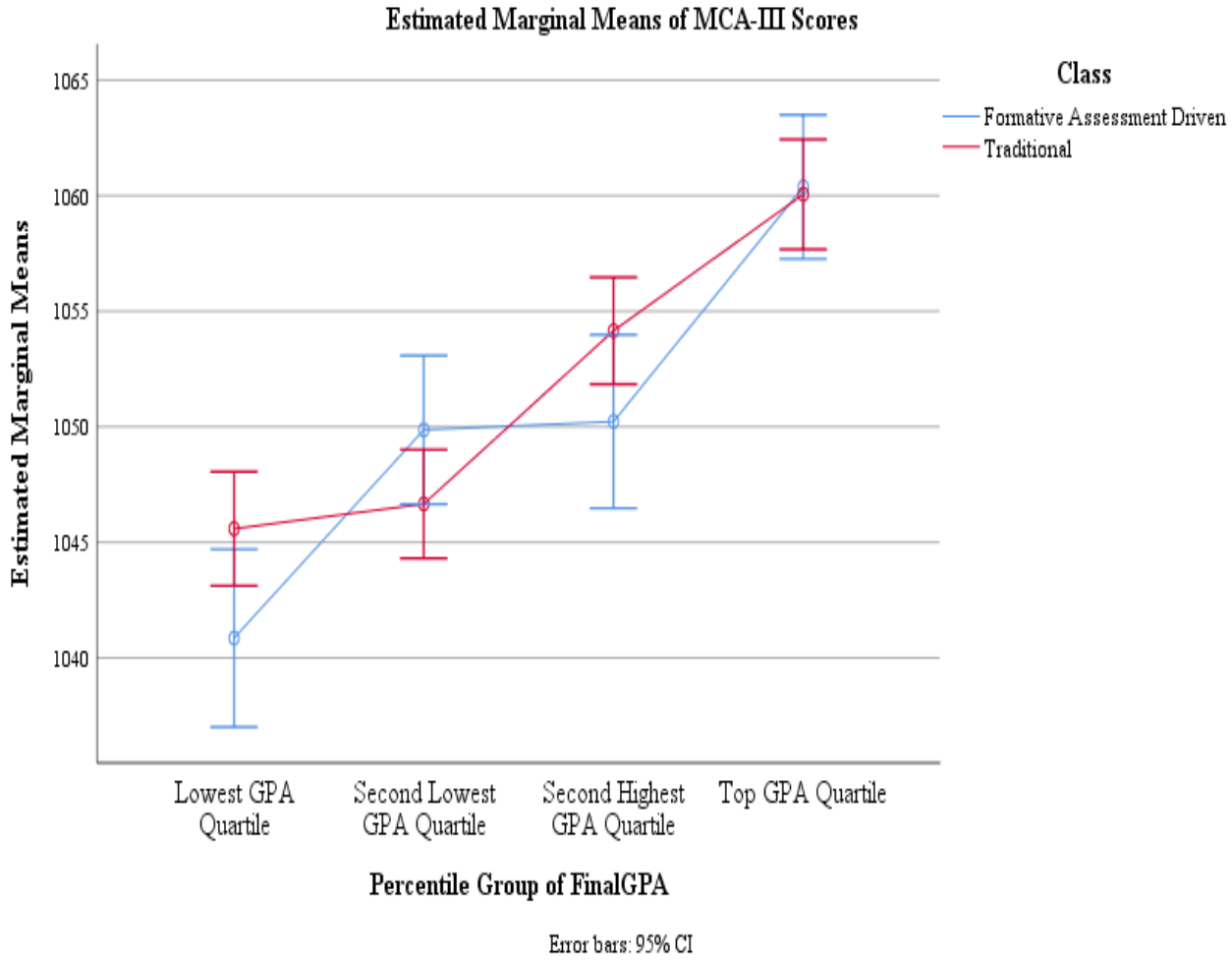
Tukey's HSD Post Hoc Tests for 2014 MCA-III Scores by GPA Quartiles

(I) Percentile Group of Final GPA	(J) Percentile Group of Final GPA	Mean Difference (I-J)	Standard Error	Sig.
Lowest GPA Quartile	Second Lowest GPA Quartile	-3.57	1.430	.062
	Second Highest GPA Quartile	-8.87*	1.455	.000
	Top GPA Quartile	-15.96*	1.427	.000
Second Lowest GPA Quartile	Lowest GPA Quartile	3.57	1.430	.062
	Second Highest GPA Quartile	-5.30*	1.391	.001
	Top GPA Quartile	-12.39*	1.362	.000
Second Highest GPA Quartile	Lowest GPA Quartile	8.87*	1.455	.000
	Second Lowest GPA Quartile	5.30*	1.391	.001
	Top GPA Quartile	-7.10*	1.387	.000
Top GPA Quartile	Lowest GPA Quartile	15.96*	1.427	.000
	Second Lowest GPA Quartile	12.39*	1.362	.000
	Second Highest GPA Quartile	7.10*	1.387	.000

Based on observed means.
The error term is Mean Square (Error) = 80.179.
*. The mean difference is significant at the .05 level.

Figure 4

2014 MCA-III Scores by Class Type and GPA Quartiles



2015 (MCA-III)

As was the case with the 2011 results, there was a significant main effect for type of instruction, F (1,320) = 4.91, p = 0.027, η^2 = .015. However, unlike the 2011 data, in 2015 students in the traditional teaching class (M = 1053.74, SD = 11.78) performed better than the students in the formative assessment driven class (M = 1050.69, SD = 12.72). There was also a significant main effect for GPA, F (3,320) = 63.14, $p < .001$, η^2 = .372. (See Table 14 for full ANOVA table.) Tukey post hoc tests revealed that all groups were significantly different from each other (see Table 13 for means and standard deviations and Table 15 for post hoc results). Similar to results in 2011-2013, in the 2015 data there was no significant interaction between type of instruction and GPA level on MCA-III scores, F(3,320) = 1.96, p = .12. However, an examination of the graph in Figure 5 does reveal that students in the lowest GPA group performed better if they were in the traditional teaching classroom.

Table 13

2015 Means and Standard Deviations for MCA-III Scores by Class Type and GPA Quartiles

Class	Percentile Group of Final GPA	Mean	Standard Deviation	N
Formative Assessment Driven	Lowest GPA Quartile	1037.88	9.731	25
	Second lowest GPA Quartile	1048.07	10.569	28
	Second highest GPA Quartile	1053.94	9.890	34
	Top GPA Quartile	1063.00	6.633	23
	Total	1050.69	12.715	110
Traditional	Lowest GPA Quartile	1044.81	10.019	53
	Second lowest GPA Quartile	1050.64	9.422	53
	Second highest GPA Quartile	1053.98	7.237	47
	Top GPA Quartile	1063.38	10.501	65
	Total	1053.74	11.775	218
Total	Lowest GPA Quartile	1042.59	10.388	78
	Second lowest GPA Quartile	1049.75	9.844	81
	Second highest GPA Quartile	1053.96	8.394	81
	Top GPA Quartile	1063.28	9.606	88
	Total	1052.72	12.165	328

Table 14

2 (Class) X 4 (GPA Quartiles) ANOVA Table for 2015 MCA-III Scores

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	19603.150[a]	7	2800.450	31.130	.000
Class	441.599	1	441.599	4.909	.027
GPA Quartiles	17040.954	3	5680.318	63.143	.000
Class * GPA Quartiles	528.429	3	176.143	1.958	.120
Error	28787.045	320	89.960		
Corrected Total	48390.195	327			

a. R Squared = .405 (Adjusted R Squared = .392)

Table 15

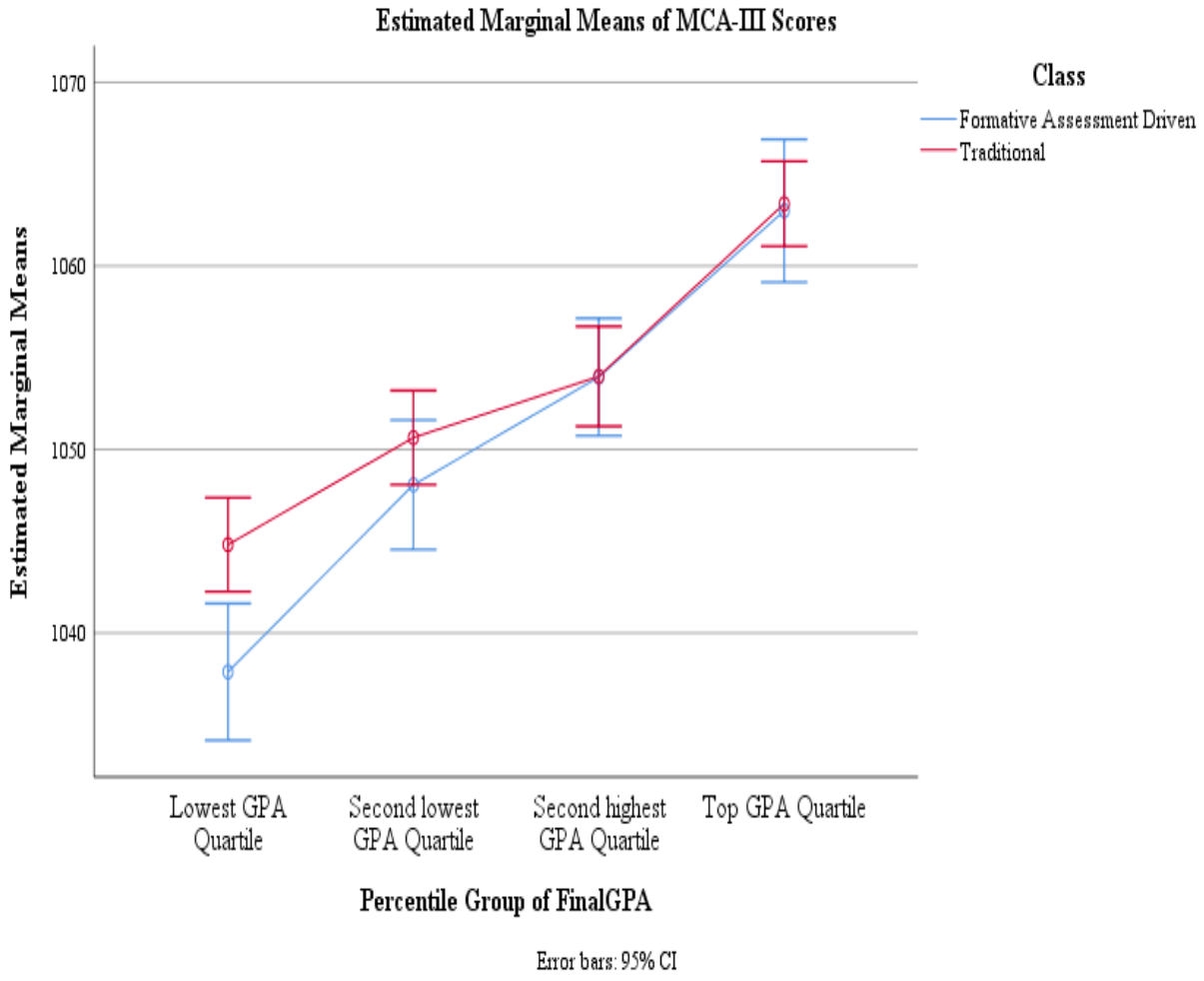
Tukey's HSD Post Hoc Tests for 2015 MCA-III Scores by GPA Quartiles

(I) Percentile Group of Final GPA	(J) Percentile Group of Final GPA	Mean Difference (I-J)	Standard Error	Sig.
Lowest GPA Quartile	Second lowest GPA Quartile	-7.16*	1.505	.000
	Second highest GPA Quartile	-11.37*	1.505	.000
	Top GPA Quartile	-20.69*	1.475	.000
Second lowest GPA Quartile	Lowest GPA Quartile	7.16*	1.505	.000
	Second highest GPA Quartile	-4.21*	1.490	.026
	Top GPA Quartile	-13.53*	1.460	.000
Second highest GPA Quartile	Lowest GPA Quartile	11.37*	1.505	.000
	Second lowest GPA Quartile	4.21*	1.490	.026
	Top GPA Quartile	-9.32*	1.460	.000
Top GPA Quartile	Lowest GPA Quartile	20.69*	1.475	.000
	Second lowest GPA Quartile	13.53*	1.460	.000
	Second highest GPA Quartile	9.32*	1.460	.000

*. The mean difference is significant at the .05 level.

Figure 5

2015 MCA-III Scores by Class Type and GPA Quartiles



Summary of Results

There was a significant result for the type of instruction used in 2011 and again in 2015. In the first year of the study students receiving formative-assessment driven instruction in the experimental group performed better on the Minnesota Comprehensive Assessment for Science II than those receiving traditional instruction in the control group. There was no significant result for type of instruction used in the next three years of the study from 2012 – 2014. The significant interaction in 2015 was the reverse of 2011 with students receiving traditional instruction performing better than those receiving formative-assessment driven instruction.

There was a consistent significant result in each of the five years of data between grade point average and performance on the Minnesota Comprehensive Assessment for Science II and III. Students in each grade point quartile, moving from lowest to highest, generally performed progressively better on the exam.

The was no significant result between the experimental and control groups for interaction between grade point quartiles, type of instruction, and performance on the Minnesota Comprehensive Assessment for Science I and II in four of the five years of this study.

Chapter V: Summary

Introduction

The United States has been in the process of adapting and restructuring its education system to meet the changing realities of the twenty-first century globalized world. The globalized era requires a more broadly educated populace as emerging economic realities have literally restructured the makeup of the workplace. An extensive standardized testing system has been developed and employed to both monitor and guide the needed educational changes. The testing system is by necessity driven by technology resulting in bulk quantitative data produced for analysis. Instructional practice has seen significant development during the globalized era educational reform efforts. Formative driven-assessment instruction has consistently produced improved student achievement.

Overview of Study

The purpose of this study was to determine whether or not a connection can be established between a decentralized formative driven-assessment approach and success in a large-scale standardized testing system. The former had significant research support while the latter was pragmatically necessary. The objective nature of a standardized testing system contradicted the individualization of the formative driven-assessment process. If significant interaction between a classroom based instructional practice and a large scale data driven assessment structure could be established a new strategy helpful on both local and state/national levels would emerge. If formative assessment-driven teaching could be linked to improved test results, a worthy path to pursue accountability would be established.

Research Question

The research question for this study was: what is the relationship between formative assessment-driven instruction and standardized test scores, particularly for average or below average students?

Hypotheses

There were three hypotheses and three null hypotheses proposed in this study:

Hypothesis One: There will be significant differences in performance on the Minnesota Comprehensive Assessment for Science scores between students receiving formative assessment-driven instruction and students receiving traditional instruction.

Null Hypothesis One: There will be no differences in the Minnesota Comprehensive Assessment for Science scores between students receiving formative assessment-driven instruction and students receiving traditional instruction.

Hypothesis Two: Student Minnesota Comprehensive Assessment for Science scores will correlate with their overall student achievement level as measured by quartiles of four-year grade point averages.

Null Hypothesis Two: Student Minnesota Comprehensive Assessment for Science scores will not correlate with their overall student achievement level as measured by quartiles of four-year grade point averages.

Hypothesis Three: There will be a significant interaction between receiving type of formative assessment-driven instruction and student achievement quartiles.

Null Hypothesis Three: There will be no significant interaction between type of formative assessment-driven instruction and student achievement quartiles.

Analysis

There was not a consistent significant result over the five years of this study regarding hypothesis one. There was a significant result in years one and five. There was not a significant result in years two, three, and four. The significant result (F (1,405) = 5.43, p = .02, η^2 = .013) from year one of the study showed students receiving formative-driven assessment instruction performing better than those receiving traditional instruction. The experimental group students outperformed the control students in all four student achievement quartiles. This was the only year of the study where this would occur.

Years two through four of the study showed no significant result between students receiving formative assessment-driven instruction and their classmates receiving traditional instruction. It is noteworthy that students receiving traditional instruction outperformed those who received formative assessment-driven instruction in the second lowest grade point average quadrant in 2012, in all quadrants in 2013, and in the lowest and second highest quadrants in 2014. Year five of the study once again showed a significant result. The result (F (1,320) = 4.91, p = .027, η^2 = .015) was unlike year one of the study in that students receiving traditional instruction outperformed those who received formative assessment-driven instruction.

The results for hypothesis two showed a significant main effect for GPA and Minnesota Comprehensive Assessment for Science results. Students in each grade point quartile, moving from lowest to highest, generally performed progressively better on the exam. Each quartile group scored significantly from each other in a linear fashion in all five years of the study with the exception of the 2nd and 3rd quartiles in 2012 and the lowest quartile in 2014. Hypothesis two was confirmed.

Hypothesis three investigated the impact of the type of instruction for students of varying achievement levels on Minnesota Comprehensive Assessment for Science results. There was no significant interaction between the independent variables for years one through three of the study. There was significant interaction (F (3,317) = 3.01, p = .03, η^2 = .028) in year four of the study. While there was significant interaction, it was not consistent. In the lowest and second highest GPA groups the traditional instruction group performed better. In the second lowest GPA group the formative assessment group performed better. In the highest GPA group there were similar scores between traditional and formative assessment groups. While the interaction from year five did not show significant interaction between the independent variables, students receiving traditional instruction performed better than those receiving formative assessment-driven instruction in the lower two quadrants of student grade point averages. The 2015 results were almost a reversal of the first year of the study with formative assessment-driven and traditional approaches exchanging positions.

Conclusion

This study produced a mix of results. Year one of the study produced data that reinforced the driving component of this study: formative assessment-driven instruction. Students receiving that form of instruction achieved better scores on the Minnesota Comprehensive Assessment for Science than students receiving traditional instruction in each of the four grade point average quadrants. That was the only time that happened over the five years of the study.

The case can be made that the formative assessment-driven instruction students fared progressively worse in each year of the study than those receiving traditional instruction culminating in year five where the traditionally instructed control group performed better in each of the four grade point average quadrants. This raises the question of how influential a work

seminal to the high school was. Was there an institutional osmosis where the traditional teachers gradually employed formative practices within a traditional structure? The highly collaborative nature of the Science department would seem to enhance the possibility that this occurred.

The results of this study also raised questions of teacher efficacy. While the experimental group had the same teacher for each of the five years of the study, the control group saw turnover during the five years of the study. Either four or five teachers taught biology each of the five years of the study. Nine different teachers were involved. Differentiating data among them could have revealed a wider and more complex rationale for the results.

Another question raised by the results lies in the fluctuating results of 2014. Why did the lower quadrant of students perform so much better in the traditional group than their formative assessment-driven counterparts? Why did this reverse itself in the 2nd lowest quadrant and then again in the 2nd highest? With no clear pattern emerging the question arises of what other factors could have caused this pattern.

Another potential question for investigation would consider the development of critical thinking and the types of assessment it could influence. A formative assessment-driven type of instruction naturally fosters critical thinking with its student centered approach. Critical thinking skills are not necessarily applicable to standardized tests given the technological structure that feature multiple choice test questions.

It is clear this study did not produce results that formative assessment-driven instruction had a significant impact on Minnesota Comprehensive Assessment for Science scores. The components that framed this study remain intact. Formative assessment-driven instruction maintains its status as a successful research backed pedagogical approach. Standardized testing remains the primary accountability tool for the foreseeable future in U.S. public schools. The

opportunity afforded in this study did not produce a significant interactive relationship between the two.

References

Bailey, K. & Jakcic, C. (2012). *Common formative assessment: A toolkit for professional learning communities at work.* Bloomington, IN: Solution Tree Press.

Barth, P. & Mitchell, R. (2006). *Standardized tests and their impact on schooling: Questions and answers*. Retrieved from https://sites.psu.edu/martinoci/2013/02/17/the-positives-of-standardized-testing/

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice*, *5*(1), 1-3, 7-74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 1-13.

Bradley, R.H., & Corwyn, R.F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, *53*, 371-399, Retrieved from http://dx.doi.org/10.1146/annurev.psych.53.100901.135233

Benjamin, S. (2014). Shifting from data to evidence for decision making, *Phi Delta Kappan*, *95*(7), 45-49.

Bill and Melinda Gates Foundation. (2010). *Primary sources: America's teachers on America's schools*, Retrieved from http://mediaroom.scholastic.com/taxonomy/term/582

Bloom, B. (1968). *Learning for mastery*. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. Retrieved from https://www.scirp.org/(S(i43dyn45teexjx455qlt3d2q))/reference/ReferencesPapers.aspx?ReferenceID=1223917

Chappuis, J. (2014). *Seven strategies of assessment for learning.* (2nd edition). Upper Saddle River, NJ: Pearson Education.

Chappuis, J. (2014) Thoughtful assessment with the learner in mind. *Education Leadership*, *71*(6), 20-26.

Christensen, C. (2008). *Disrupting class: How innovation will change the way the world learns*. NY: McGraw Hill.

Clark, I. (2011). Formative assessment: Policy, perspectives and practice. *Florida Journal of Educational Administration and Policy*. *4*(2), 158-180.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). Retrieved from http://olam.ed.asu.edu/epaa/v8n1/

Duncan, G. & Murname, R. (2014). Meeting the educational challenge of income inequality, *Phi Delta Kappan*, *95*(7), 50-54.

Druckor, B. (2014). Formative assessment in seven good moves, *Education Leadership*, *71*(6), 28-32.

Elgart, M. (2017). *Meeting the promise of continuous insights from the advancED continuous improvement system and observations of effective schools.* Alpharetta, GA: Advance Education Inc. Retrieved from www.advanceded.org/sites/default/files/CISWhitePaper.pdf

Every Student Succeeds Act, PL 114-95, 20 U.S.C. 6301 (2015)

Ferguson, M. (2016). ESSA is more than the latest acronym on education's block. *Phi Delta Kappan*, *97*(6), 72-73.

Friedman, T. (2005). *The world is flat; a brief history of the 21st century*. NY: Farrar, Straus, and Giroux.

Frey, B.B. & Schmitt, V.L. (2010). Teachers' classroom assessment practices. *Middle Grades Research Journal*, *5*(3), 107-117.

Gallup and NWEA. (2016). Make assessment work for all students: Multiple measures matter. *Phi Delta Kappan*, *99*(1), 22-34.

Goe, L., & Stickler, L. (2008). Teacher quality and student achievement: Making the most of recent research. *National Comprehensive Center for Teacher Quality*, Washington, D.C.

Gordon, B. (2007). U.S. competitiveness: The education imperative. *Issues in Science and Technology*, *23*(3)1-12.

Guskey, T. (2019). Grades versus comments: Research on student feedback. *Phi Delta Kappan, 101*(3), 42-47.

Gutek, G. (2011). *Historical and philosophical foundations of education: A biographical introduction.* Upper Saddle River, NJ: Pearson.

Hattie, J. (2015, February). High impact leadership. *Educational Leadership*, *72*(5), 37-40.

Hattie, J. (2015, October). The effective use of testing: What the research says. *Education Week*, *80*(28), 23.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* NY: Routledge.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Hauushek, E. (2014). Why the U.S. results on PISA matter. *Education Week*, *33*(15), 20.

Huxham, M. (2007). Fast and effective feedback: Are model answers the answer? *Assessment and Evaluation in Higher Education, 32*(6), 601-611.

Jennings, J. (2018). It's time to redefine the federal role in K-12 education. *Phi Delta Kappan*, *100*(1), 7-11.

Jouriles, G. (2014). We don't need standardized tests. Here's why. *Education Week*, *33*(36), 36-40.

Kamaenetz, A. (2018). What 'a nation at risk got wrong, and right, about U.S. schools. *National Public Radio: How Learning Happens.* Retrieved from: https://www.npr.org/sections/ed/2018/04/29/604986823/what-a-nation-at-risk-got-wrong-and-right-about-u-s-schools

Koenka, A., & Anderman, E. (2019). Personalized feedback as a strategy for improving motivation and performance among middle school students. *Middle School Journal*, *50* (5), 15-22.

Koretz, D., & Hamilton, L. (2006). *Educational measurement*. Westport, CT: American Council on Education, p. 531-578

Koretz, D. (2017). *The testing charade: pretending to make school better*. Chicago, IL: University of Chicago Press.

Kulhavy, R.W. (1977). Feedback in written instruction. *Review of Educational Research*, *47*(1), 211-232.

Leachman, M., Masterson K., & Figueroa, E. (2017). A punishing decade for school funding. *Center on Budget and Policy Priorities*, Retrieved from https://www.cbpp.org/research/state-budget-and-tax/a-punishing-decade-for-school-funding

Locke, E.A., & Latham, G.P. (2002). Building a practically useful theory of goal setting and task motivation: A 35 year odyssey. *American Psychologist*, *57*, 705-717.

Loomis, S.C., & Bourque, M.L. (2001). *National Assessment of Educational Progress achievement levels, 1992-1998 for reading.* Washington D.C.: National Assessment Governing Board.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R.J. (2010). *Formative assessment & standards based grading.* Bloomington, IN: Marzano Research Laboratory.

McNeil, M. (2011). More states are asking for NCLB waivers. *Education Week*, *30* (37), 20.

McTighe, J. (2018). Three key questions on measuring learning. *Educational Leadership, 75* (5) 14-20.

Mehta, J. (2013). *The allure of order: High hopes, dashed expectations, and the troubled quest to remake American schooling.* Oxford: Oxford University Press.

Moss, C.M., & Brookhart, S.M. (2010). Teacher inquiry into formative assessment practices in remedial reading classrooms. *Assessment in Education: Principles, Policy & Practice, 17*(1), 41-58.

Mullis, I.V.S., Martin M.O., Foy, P., & Drucker, K.T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

National Center on Education and the Economy. (1990). *America's choice: High skills or low wages!* Retrieved from http://ncee.org/wp-content/uploads/2013/09/Americas-Choice-High-Skills-or-Low-Wages.pdf

No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. 6319 (2002).

Phelps, R. (2002). *Estimating the costs and benefits of educational testing programs.* Education Consumers Foundation. Retrieved from http://www.education-consumers.com/briefs/phelps2.shtm

Phelps, R. (2011, April). The effect of testing on achievement: Meta-analyses and research summary, 1910–2010, *Nonpartisan Education Review.*

Piaget, J. (1970). *Genetic epistemology*. NY: Columbia University Press.

Popham, J. (2006). *Defining and enhancing formative assessment.* Washington D.C.: Council of Chief of State School Officers.

Popham, J. (2013). Waving the flag for formative assessment. *Education Week*, *32*(15) 29.

Ravitch, D. (2010). *The life and death of the great American school system: How testing and choice are undermining education.* NY: Basic Books.

Rudalevige, A. (2003). The politics of no child left behind. *Education Next*, *3*, 63-69.

Sadler, D.R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 2.

Schneider, J., Feldman, J., & French, D. (2016). The best of both worlds. *Phi Delta Kappan.* Retrieved from https://kappanonline.org/schneider-feldman-french-grading-standardized-tests-best-both-worlds/

Scriven, M. (1967). *The methodology on evaluation*: *Perspectives on curriculum*. Chicago: Rand McNally and Company.

Schwab, K. (2016). *The global competition report 2016-2017.* Geneva: Global Economic Forum.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Shute, V.J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153-189.

Stiggins, R. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement Issues and Practice*, *20*(3), 5-15.

Stronge, J. H. (2002). *Qualities of effective teachers*. Alexandria, VA: Association for Supervision and Curriculum Development.

Superville, D. (2015). Study paints a chaotic view of testing. *Education Week*, *35*(10), 1, 9.

Tanis, B. (2014). Pushing back against high stakes for students with disabilities. *American Educator*, *38*(4), 20.

The 49th Annual Phi Delta Kappan poll of the public's attitude towards the public schools: Academic achievement is not the only mission. *Phi Delta Kappan*, *99* (1) 25-27.

Tofler, A. (1990). *Powershift: Knowledge, wealth, and violence at the edge of the 21st century.* NY: Bantam Books.

Tomlinson, C. (2014). The bridge between today's lesson and tomorrow's. *Education Leadership*, *71*(6), 10-14.

Tucker, M. (2015). Needed: an updated accountability model. *Educational Leadership*, *72*(5), 66-70.

Tucker, M., & Codding, J. (2002). *Standards for Our Schools.* San Francisco: Jossey-Bass.

U.S. Department of Education. (1999). *Taking responsibility for ending social promotion. A guide for educators and state and local leaders.* Retrieved from https://files.eric.ed.gov/fulltext/ED430319.pdf

U.S. Department of Education. (2004). *Testing: Frequently asked questions*, Retrieved from https://www2.ed.gov/nclb/accountability/ayp/testing-faq.html

U.S. Department of Education, National Center for Educational Statistics. (2012). *NCES statement on PIRLS and TIMSS 2011 results.* Retrieved from https://nces.ed.gov/timss/results11.asp

U.S. Department of Education, National Center for Educational Statistics. (2019). *School choice in the United States: 2019*. Retrieved from https://nces.ed.gov/programs/schoolchoice/

U.S. Department of Education, National Center for Educational Statistics. (2019). *The nation's report card: Reading 2011*, Retrieved from https://nces.ed.gov/nationsreportcard/pdf/main2011/2012457.pdf

U.S. Department of Education, National Center for Educational Statistics. (2016). *TIMSS and TIMSS advanced 2015 results.* Retrieved from https://nces.ed.gov/timss/timss15advanced.asp

U.S. Department of Education, National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Retrieved from https://www.edreform.com/wp-content/uploads/2013/02/A_Nation_At_Risk_1983.pdf

Views of a changing landscape, (2014). *Education Week*, 33(16), 10-11, 20-21, 24.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wang, L., Gulbahar, H. B., & Brown, L. (2006). Controversies of standardized assessment in school accountability reform: A critical synthesis of multidisciplinary research evidence. *Applied Measurement in Education*. *19*(4) 305-328.

White, G., Stepney, C., Hatchimonji, D., Moceri, D., Linsky, A., Reyes-Portillo, J., & Elias, M. (2016). The increasing impact of socioeconomics and race on standardized academic test scores across elementary, middle, and high school, *American Journal of Orthopsychiatry*, *86* (1) 10-23.

Whitehurst, G. (2014). *The future of test-based accountability*. Brookings Institute, Retrieved from https://www.brookings.edu/research/the-future-of-test-based-accountability/

Wilburn, G., Cramer, B., & Walton, E. (2020). *The great divergence: Growing disparities between the nation's highest and lowest achievers in NAEP mathematics and reading between 2009 and 2019*. Washington D.C.: U.S. Department of Education, National Center for Educational Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/blog/mathematics_reading_2019.aspx

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment*, *11,* (3), 283-289,

Wiliam, D. (2007). Changing classroom practice. *Education Leadership*, *65*(43), 36-42.

Wiliam, D. (2014). The right questions, the right way. *Education Leadership*, *71*(6), 16-19.

Withers, M. (2014). Seeing opportunities in 21st century challenges, *Education Week*, *33*(16), 32-33.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57–67.

Yeager, D. and Dweck C. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, *47*(4), 302–314.

Yeh, S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy Analysis Archives, 13,* 43. Retrieved from https://epaa.asu.edu/ojs/article/view/148/274