2012-01-01

Distortion, Disparity, and Dubious Data: The Impact of Accountability on Instructional Practice

Curtis Jack Barnes
University of Texas at El Paso, cbarnes003@elp.rr.com

DISTORTION, DISPARITY, AND DUBIOUS DATA:

THE IMPACT OF ACCOUNTABILITY ON INSTRUCTIONAL PRACTICE

CURTIS J. BARNES

Department of Educational Leadership and Administration

APPROVED:

_____
Don Schulte, Ed.D., Chair

_____
Bill Johnston, Ph.D.

_____
Penelope Espinoza, Ph.D.

_____
George Keating, Ed.D.

_____
Benjamin C. Flores, Ph.D.
Interim Dean of the Graduate School

DISTORTION, DISPARITY, AND DUBIOUS DATA:

THE IMPACT OF ACCOUNTABILITY ON INSTRUCTIONAL PRACTICE

By

CURTIS J. BARNES, B.A., M.F.A., M.Ed.

DISSERTATION

Presented to the Faculty of the Graduate School of

the University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF EDUCATION

The Department of Educational Leadership and Administration

THE UNIVERSITY OF TEXAS AT EL PASO

May 2012

ACKNOWLEDGEMENTS

Completion of this dissertation, let alone the doctoral program which preceded it, would not have been possible without the unwavering support of scores of individuals too numerous to name here. Nevertheless, I shall try to mention a few who played prominent roles in helping me through to this accomplishment, with sincere apologies to anyone whom I may miss.

First, I cannot find sufficient words to express my gratitude to Dr. Don Schulte who not only supported me through this process with thoughtful advice and insightful feedback, but who also displayed remarkable patience and understanding, certainly beyond what was warranted, as well as provided a firm but gentle push when necessary. Without his guidance and wisdom, this study would surely have remained a loose collection of ideas seeking a direction.

Special thanks is also due to my committee members each of whom provided invaluable contributions to this effort. Dr. Bill Johnston, Dr. Penelope Espinoza, and Dr. George Keating all shared knowledge and expertise critical to allowing these ideas to take shape. Without their combined efforts it is difficult to imagine how this work would have come to fruition.

In addition, I would like to mention Dr. Mike Warmack, who inspired me to enter this program and whom I have sometimes callously blamed for many lost hours of sleep. You were right, Mike; it was a good thing. In addition, I should not fail to thank my former boss in Research and Evaluation, Art Jordan, who always supported this effort, as well as my current leader, Dr. James Steinhauser, who has likewise been a source of continued encouragement and support. Several of my other colleagues in EPISD, including Dr. Nidelia Montoya, who helped me make sense of the TEA data files, Anna Bone, Joe Hernandez, Yinou Duo, Carlos Perales, Ruby Lynch Arroyo, and Kim Arispe, each offered important support that facilitated this study.

To my friend Bill Anderson, who is like family, and who endured hours of conversation about school accountability and managed to pretend interest, I owe you, bud. Most importantly,

to my family, who put up with countless late dinners and piles of journal articles, books, and data tables, there is no way to adequately say thank you. I can only hope that my acknowledgement that none of this would have been remotely possible were it not for the understanding of my wife Maria, especially, and my two sons, Jack and Matthew, will say it for me. Without my parents, Jack and Rilla Barnes, I would not have succeeded in this program nor in anything else. To the extent that this is my accomplishment, it is also theirs.

THE UNIVERSITY OF TEXAS AT EL PASO

ABSTRACT

DISTORTION, DISPARITY, AND DUBIOUS DATA: THE IMPACT OF ACCOUNTABILITY ON INSTRUCTIONAL PRACTICE

This study examined the impact that state and federal accountability systems have had on instructional practice in two large Texas school districts by comparing the performance of students at these schools on individual items from the 2011 Texas Assessment of Knowledge and Skills (TAKS) and relating performance to item difficulty and the schools' accountability risk as determined by prior accountability performance. To make this comparison, schools were placed into accountability risk groups based on past performance on the No Child Left Behind Act's (NCLB) Adequate Yearly Progress (AYP) accountability instrument. The researcher then calculated the mean differences between average performance on each item and compared them against the item difficulty score to determine if the relationship was significant. The relationship was tested for all groups, both isolating and excluding economically disadvantaged students and those with limited English proficiency. Inclusion in these groups was based on the coding the students were assigned in the state's assessment data file. This comparison used a simple regression analysis performed on SPSS statistical software, version 20.

The findings of the study revealed a statistically significant positive relationship between the gap in performance for each risk group and the item difficulty level, meaning that as item difficulty increased the gap between students in the various risk groups grew larger as well. Based on this study, the researcher questions how effective the federal accountability system is at determining school achievement.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Background

In January 2002, President Bush signed into law the *No Child Left Behind Act of 2001* (NCLB), reauthorizing the *Elementary and Secondary Education Act of 1965* and ushering in a new era of educational accountability. Far from a seminal moment, however, NCLB actually represented a crescendo of sorts–a *fait accompli* marking the culmination of a gradual shift toward outcome-based accountability in education that had been gathering momentum since the issuance nearly two decades earlier of *A Nation at Risk: The Imperative for Educational Reform*.

The law required all 50 states, as well as Puerto Rico and the District of Columbia (DC), to develop performance-based accountability systems by 2006. The key element of the system was its ambitious requirement that states demonstrate that 100 percent of students achieve proficiency in reading and mathematics by 2014. To ensure compliance, the law predicated Title I funding on states implementing annual measurable objectives (AMOs) in both mathematics and reading that would apply to districts, schools, and designated subgroups within the schools that met prescribed size requirements. States were required to set interim benchmarks to demonstrate adequate yearly progress (AYP) toward the law's 100 percent proficiency goals, with schools and districts failing to achieve AYP subject to a series of escalating sanctions that would potentially lead to school restructuring requiring all or most of the school's staff to be replaced, including the school principal.

NCLB as well as the state accountability systems which preceded it and from which it borrowed many of its provisions, is based on the premise that the accountability system's requirements and potential sanctions would act as an incentive for educators to improve

instruction and thereby increase learning and achievement for students. The acceptance of this premise was wide-spread even prior to the codification of it into law but has of course grown under NCLB. More than half of the states now utilize performance criteria on state assessments as a factor in promotion and/or graduation. Among those states 18 use some form of financial incentive or reward to teachers and administrators related to state accountability as primarily determined by standardized assessments. As of 2008, at least 32 state accountability systems, including Texas's Academic Excellence Indicator System (AEIS), provide for potential sanctioning of school staff on the basis of poor student performance (Center on Education Policy, 2006). Moreover, with the enactment of NCLB, public educators nationwide are subject to sanctioning provisions of AYP, the federal accountability requirements.

In terms not only of scope but also with regards to impact, accountability mandates dwarf all other education policy reforms. In fact, more than 90 percent of the nearly 53 million children attending elementary and secondary schools in the United States are enrolled in public schools subject to federal accountability under NCLB (School Data Direct, 2009).

Proponents of test-based accountability have long argued that accountability creates an incentive for students, parents, teachers, and administrators to work harder and forces educators to focus on identifying struggling students and schools and their educational needs (Bishop & Mane, 2001; Ravitch, 1996; Stotsky, 2000). In theory, these factors will improve student performance by raising motivation, increasing parent involvement, and encouraging states and districts to improve curriculum and pedagogy (Simmons & Resnick, 1993; Spalding, 2000; Viadero, 1994).

However, as researchers have long noted, major policy initiatives are perceived and implemented through a variety of highly disparate lenses, and thereby often result in substantial

unintended consequences. Hence, Deborah Stone (1997) postulates, in *Policy Paradox*, that implementation of policy "does not reliably follow economic models of markets and incentives" (p. 34). Further, complex economic theory has consistently demonstrated that both incentives and disincentives generally yield unforeseen and typically undesirable distortions. For example, Holmstrom and Milgrom (1991) demonstrate that incentive/disincentive devices, "though based on seemingly objective criteria, incline actors to focus on the most easily observable criteria or outcomes" (p. 29). Several examples of such distortions may be seen in the general impacts on school programs of both state and federal accountability systems, including such commonly raised concerns as narrowing curricula, sample distortion — commonly described as teaching to the test — along with a disproportionate focus on ensuring that all students attain at least minimum achievement standards to the detriment of higher-order and/or critical thinking skills. Moreover, critics such as Koretz (2000) and Linn (2009) note that such accountability policies have induced districts, schools, and teachers to divert resources from subjects, especially fine arts, that are not explicit indicators of accountability to focus on areas such as math and reading that are typically measured directly under such models as well as to neglect students well outside the margins, namely students deemed probable failures or probable passers, to concentrate on the most easily attained progress indicators (in the case of education, the so-called bubble students) where outcomes are in doubt. Not only does such distortion ignore critical aspects of the curriculum, such as social studies, that are not explicitly measured, but it also disserves both the most needy and most promising students in our schools. As Padilla (2005) has observed, the focus educational entities are placing on outcome or numbers-based accountability models such as the Texas AEIS and the federal AYP provisions of NCLB, has led to a distortion of traditional educational goals, such as passing an examination rather than mastering a skill, as it is clear that

the former does not necessarily require the latter. Preparing students to pass the state test corrupts the institution, distorting the school's mission "from one of education to one of self preservation" (Padilla, 2005, p. 256). Achieving this distorted goal often induces educators to modify instructional practices away from traditional substantive content to what Padilla dubs "heuristics of test taking," more commonly referred to as test taking skills (p. 257).

The unique pressures and challenges typically faced by lower-performing schools, including disproportionate levels of poverty, limited English proficiency, and deficient levels of parental education, only exacerbate these issues. In fact, even within communities, these challenges disparately impact some schools and school children more than others. Indeed, such disparity and distortion can even be observed within individual campuses and classrooms. Though the debate continues to rage with regard to the promise and peril incumbent with high stakes testing and accountability models which focus on such testing, evidence is increasingly clear that such models have significantly impacted and changed the school experience for American students, especially those who attend schools that face significant and unique challenges not adequately provided for within the system (Abedi, 2004; Kim & Sunderman, 2005; Tracey, Sunderman, & Orfield, 2005).

It follows, therefore, that the distortions manifested as a result of outcome/numbers-based accountability will intensify in schools in relationship to the challenges such schools encounter in their efforts to meet the increasingly demanding provisions of NCLB (the so-called, "Moving Targets) as well as similarly challenging provisions of state accountability systems.

Despite its widespread scope and impact on education, scant research is available on the disparate impacts of test-based accountability and its chief component, high-stakes standardized tests on students, teachers, and campuses/districts facing these challenges to a disproportionate

degree and the resulting distortions to the educational programs that serve them. However, the hue and cry from teachers in these institutions has never been louder. Anecdotal accounts of such distortion are abundant, but the very nature of the problem makes it difficult to reliably quantify. By contrast, advocates of NCLB and other test-based accountability systems can point to metrics that reinforce the idea that test-based accountability has led to gains, especially among minority students. Such growth is not only visible in the frequently derided and oft scorned state standards and their associated and equally denigrated assessments, but also in the much celebrated and cited National Assessment of Educational Progress (NAEP). While it is true that student proficiency levels under NAEP tend to be far below those of state measures, students have, nevertheless, shown progress in NAEP, especially minority student groups (Whitehurst, 2010).

Critics, meanwhile, are left with a cacophony of disparate voices which rely on the argument that test-based accountability has induced educators to emphasize accountability ratings at the expense of students' education (Koretz, 2005; Padilla, 2005). While such actions are often defended as the well-meaning result of educators trying to shield students from the impact of NCLB's accountability provisions, such as grade retention, policy analysts note that such concerns ignore the realities of the legislation. To the contrary, NCLB contains liberal provisions which substantially shield students from this type of direct impact. Moreover, these actions also call into question educational ethics since the strategies utilized by educational entities to meet AYP provisions frequently do just the opposite, as when, for example, students are retained in non-measured grades to minimize accountability exposure (Guggino & Brint, 2010).

One frequently cited unintended consequence of test-based accountability is that educators disproportionately target their educational program to the threshold level (Ravitch, 2010; Rothstein, 2008). Such targeting manifests itself in two distinct but related strategies: (1) instruction is primarily designed and focused to ensure that marginal students at or near minimum proficiency successfully pass the assessment utilized by the accountability instrument (e.g., TAKS in Texas); and (2) curriculum is likewise designed to focus on the minimum skills necessary to achieve such proficiency at the minimal depth and complexity necessary. In other words, educators have determined that, from an accountability standpoint at least, there is more value in directing educational resources to bring the most students possible to the minimal level, than there is in attempting to maximize achievement for each individual student. Often referred to by the ironic and wholly inappropriate moniker of "cream-skimming," such practices result, as earlier mentioned, in the effective neglect of students at the higher and lower bounds (Finn & Ravitch, 2007; Rothstein, 2008). Students who are low enough to be deemed unlikely to pass regardless of interventions may not receive intervention suitable to their needs, if they receive any at all. This inhibits growth, meaning they are likely to remain below the target level on a consistent basis. On the other end of the scale, students deemed likely to pass may not be challenged with curricular rigor sufficient to advance them to the extent of their potential (Jacob, 2005).

This study sought to determine whether evidence could be found to support the idea that high-stakes accountability results in schools, whether intentionally or unintentionally, adopting practices that direct resources and time disproportionately toward marginal students and/or at the minimum skills level. Such practices are difficult to discern from the school performance data utilized for accountability and which is the subject of media reports and press releases that focus

attention on passing rates and school ratings but seldom explore whether schools and districts are being successful in moving students beyond minimum expectation levels, though recent efforts to measure college readiness do show some promise in this area.

Understandably, as a form of educational triage, the tendency is to worry first about schools' abilities to deliver a minimum standard of education and about students acquiring at least such, and second about moving beyond that minimum standard. However, the long-term effects of prioritizing certain students over other students (both higher and lower performing) with a complementary but restrictive curricular model can have long-term and very serious ramifications not only for these students, but also for the communities where they live and the competiveness of the nation at large (Dee, 2010).

This study makes several assumptions as to how such distortion may be impacting instruction as well as deductions regarding what may logically result there from, namely: (1) if high-stakes accountability is creating a distorted or coercive environment for lesson design and delivery, it would follow that schools that are more exposed to accountability risks, namely low- and marginal-performing schools in jeopardy of missing AYP, will show more distortion/coercion; and (2) the greater the accountability risk, particularly the danger of missing AYP, the greater the impetus toward distortion/coercion will be. By extension, this study postulates that such distortion/coercion will manifest itself in state assessments at the individual item level, rather than in percentages of students meeting minimal standards. In other words, though schools may have similar "passing" rates on the state assessment, schools with less distortion in theory would have more liberty to focus on higher-level skills and to challenge higher-performing students to advance. This study posits that these schools will show

significantly higher performance on the most difficult items of the assessment, to reflect such liberty, though the schools may have very similar proficiency rates.

Statement of the Problem

NCLB and similar accountability systems are based on the premise that holding schools accountable for student performance will work as an incentive for schools (and school districts), as well as teachers and administrators, to improve instructional practices so as to maximize student achievement. However, recent studies seeking to determine whether this premise is being realized in practice have been inconclusive, with some studies finding improvements (Braun, 2004; Carnoy and Loeb, 2002; Hanusheck and Raymond, 2004) while other studies describe negligible impact on such achievement (Amrein and Berliner, 2002; Nichols, Glass, Berliner 2006). Carnoy and Loeb (2002) adopted an index (0-5) as defined by the Consortium for Policy Research in Education (CPRE) to determine the degree of external pressure exerted by the accountability instruments to which schools and districts were subjected to as a means to improve student achievement. The index weighs a number of factors such as the application of statewide standards, the presence of sanctions/rewards linked to assessment performance, and whether or not grade retention/promotion and/or graduation were linked to such performance. Carnoy and Loeb (2002) found that the greater the external pressure (deemed strength of the accountability system), the greater the improvements in student achievement as measured on 4th and 8th grade mathematics tests. The study also found a positive correlation with regard to high school retention rates.

Hanushek and Raymond (2003, 2004) also found that students in states that had implemented stringent accountability measures in the interim between NAEP administrations showed more improvement than students in states with no accountability system on the

following NAEP administration. The study looked at students' NAEP test scale scores in reading and mathematics. Further, student achievement gains were found not only in the overall testing cohorts, but in socioeconomic subgroups as well, including African American and Hispanic groups. However, an analysis of the gap between White students' performance and that of African Americans and Hispanic students was mixed, with the gap narrowing slightly between White and Hispanic students while increasing slightly between White and African American students.

Other studies, however, have painted a far less encouraging picture as to the impact that such accountability systems have had on student achievement. For example, studies conducted by Amrein and Berliner (2002) failed to identify consistent patterns of student achievement associated with accountability when independent assessments were used to calculate achievement. Their initial study analyzed assessment data from eighteen states seeking to quantify the effect of high stakes accountability on student achievement. Amrein and Berliner concluded that since states can and do manipulate individual assessment criteria and because state assessments vary widely from state to state, a uniform, independent instrument must be employed to measure achievement (and improvement, if any). Their study analyzed test results on the ACT, SAT, NAEP, and AP, assessments, looking for a correlation between stringent accountability measures and increases in student achievement.

Utilizing the uncertainty principle, the concept that precise simultaneous measurement of some complementary variables is impossible, Amrein and Berliner came to the conclusion that there was no clear evidence of improved student learning regardless of whether the students improved their scores from previous assessments. The researchers argued that even assuming the ACT, SAT, NAEP, and AP tests are reasonable measures of the states' curriculum which

accountability incentives are intended to affect, there is insufficient evidence to conclude that accountability is driving such improvement, and therefore insufficient evidence to determine that accountability instruments do in fact have the desire incentive effect. Although year over year scores on states' high-stakes tests may show increases, Amrein and Berliner (2002) posit that the "transfer of learning is not inexorably the underlying correlation to such outcomes as suggested by accountability policy" (p. 52). In fact based on their work, Amrein and Berliner have proposed a new social-sciences version of the Heisenberg Uncertainty Principle: "The more important that any quantitative social indicator becomes in social decision-making, the more likely it will be to distort and corrupt the social process it is intended to monitor" (p. 5). Amrein and Berliner (2002) contend that such distortion, which is supported by "numerous reports of unintended consequences associated with high-stakes testing policies (increased drop out rates, teachers and schools cheating on exams, teachers' defection from the profession . . .) makes clear the need for continued study and reform of accountability policies" (p. 2).

A plethora of critics now oppose the use of assessments as the primary metric for accountability purposes due to a variety of factors ranging from score volatility, assessment deviation, and the tendency of the practice to compromise instructional practice and thereby threaten the validity of the very scores upon which the instruments rely (Nichols, Glass, & Berliner, 2006). Even test measurement specialists, as Ravitch (2010) notes, generally oppose the use of standardized tests for accountability, citing a 1999 report from the Committee on Appropriate Test Use of the National Research Council wherein psychometricians warn that "tests are not infallible" (p. 153). As a leading psychometrician, Robert Linn (2009), has explained, there are a variety of reasons that students fail tests and "failing results are not necessarily indicative of the quality or lack thereof present in a school" (p. 198).

However, Amrien and Berliner's (2002) work has not gone unchallenged. Critics noted that the researchers' initial analysis of achievement trends on the NAEP, for example, wherein the researchers compared the performance of K-8 and high school students against the national average and then organized them by state into groups that exhibited either "strong" or "weak" evidence of increases or decreases related to accountability requirements, and for which the researchers found no consistent effect related to accountability policy, was flawed because the study did not include a control group. After adding a control group and correcting what were viewed as additional method and design flaws, Rosenshine (2003) concluded that average NAEP increases were indeed greater in states with stringent accountability models than in states with less stringent models, while nevertheless conceding, after disaggregating data by state, that stringent accountability policy was "not effective policy in all states" (p. 4).

After Rosenshine's (2003) analysis, Amrein-Beardsley and Berliner (2003) adapted their research methods to include a control group but also made adaptations to control for exclusion rates (i.e., the exclusion of students from NAEP testing based on school officials determination that the students could not meaningfully participate or could not meaningfully participate without accommodations not available for NAEP testing – generally learning disabled students and students with limited English proficiency). Amrein-Beardsley and Berliner (2003) concluded that the gains Rosenshine (2003) correlated to the presence of strong accountability requirements, were insignificant when results were controlled for such exclusion, a finding they argued supported the widely held belief that such accountability systems act more as an incentive to manipulate testing cohorts (i.e., exclude low performing students from testing) than as an instructional incentive to improve learning for all students.

Henry Braun (2004) also studied achievement gains on NAEP for students in relation to the strength of the relevant accountability system applied and found that when standard error measurements were included, students in states with strong accountability instruments in place, did better on math assessments than students in states with weak or no test-based accountability. However, Braun noted that when cohorts of these students were followed as they progressed through schools, these gains "largely disappeared" (p. 33), suggesting that any gains related to accountability pressures are short-lived and potentially the result of enhanced effort by students on the assessment itself rather than an indication of actual increased achievement or improved instructional practice.

Clearly, the debate as to the efficacy of accountability systems as an incentive to improve instructional practice and enhance student achievement remains unsettled. However, it is worth noting, that while these studies have sought to examine the positive effects (or lack thereof) of such systems on student achievement, they did not attempt to account for negative unintended consequences that may have played a role in the data ultimately gathered but for which the design of the studies was inadequate to measure. In other words, are the unintended, yet "corrupt" (Jones, Jones, & Hargrove, 2003, p. 158) and "perverse" (Ryan, 2004, p. 39), consequences of such accountability instruments in fact negatively impacting student achievement even while data seems to arguably suggest that, at least for some students, the instruments are positively impacting the instruction they received. Notable, especially, are the charges that, as suggested by Amrein and Berliner (2003), NCLB has in fact distorted and corrupted the educational practices it was implemented to monitor. Are educators, in practice, under serving students as the educators struggle to adapt instructional practices that maximize performance as defined by the accountability instrument? Have these efforts come at the

expense of students outside the margins? Have they created an artificial ceiling for all students as educators seek to widen the breadth of student achievement by compromising the depth?

Research Questions

Bearing in mind the context of the current educational environment and its related focus on accountability, research into the impact of accountability systems on the processes and practices of schools and school districts and the extent to which such impact is exacerbated by external factors such as the unique challenges facing schools and school districts with high populations of economically disadvantaged, immigrant, and/or limited English proficient students remains limited. Therefore, this study sought to determine whether test-based accountability instruments have a disparate impact on instructional practices related to schools' accountability risk. I posed two fundamental questions about the ways in which schools and school districts have responded to high-stakes accountability policies:

1. Does state assessment data support the theory that high stakes accountability systems encourage educators to disproportionately direct (distort) instructional practices to minimum skills levels?
2. Does the distortion of instructional practices, if any, increase subject to the accountability exposure of schools?

Theoretical Framework

Public school educators face increasing pressure to improve the achievement of all students. Even assessment advocates such as Black and Wiliam (1998) have long recognized the potential perils inherent in high-stakes, test-based, accountability. Though opposed to the single, summative measure utilized by NCLB and most state accountability systems, preferring instead some system of formative assessment and continuous feedback, they acknowledge that outcome-

based accountability systems (including NCLB) and their associated requirements have at least focused educators on outputs instead of inputs, that is, learning instead of instruction. (Black, Wiliam, et al, 2003). Black and Wiliam promote the argument that outcomes, as a measure of student learning, are the only effective indicator of educational efficacy. However, as is increasingly clear as the impacts of such accountability systems become more apparent, outcomes as measured by a quantitative metric are subject to a variety of distortions which not only call into question the validity of the ratings issued under such systems, but also negatively impact student learning, especially for those students most in need.

This research endeavor offers an analysis of organizational decision making in response to external pressures created by state and federal accountability systems, and the extent to which disparate challenges faced by some districts, campuses, and teachers distort or alter such decisions.

Social science researchers building on rational choice and incentive theories borrowed from economics have closely examined how individuals react to incentives (rewards) and disincentives (sanctions) and the extent to which such behaviors distort organizational goals. As economists Laffont and Martimort (2001) note, the use of incentives to promote a desirable action (in this case improved instructional programs and practices) creates a paradox between institutional goals and individual goals (p. 393). A key component of incentive theory is a concept economists refer to as nonverifiability and deals with the processes that an agent pursuing an incentive utilizes to achieve a measured goal. However, broad public policy goals such as those typically established in education, for example to improve student learning, are extremely complex and difficult to measure with direct, quantifiable metrics, resulting in the establishment of indirect metrics, such as proficiency rates on standardized assessments, which

tend to corrupt public service (Simon, 1978). According to Rothstein (2008), conventional measurements of such outputs are "oversimplified and unable to support valid accountability" (p. 5). While an increase in the percentage of students testing "proficient" on standardized assessments may be an indicator of improved student learning, it could also be an indicator of several alternate factors, including more motivated students, less rigorous tests, and lowered proficiency thresholds, which have little to do with improved instructional practices or increased student learning. Likewise, a decrease or static measure in the percentage of students testing proficient may or may not indicate something with respect to the efficacy of educational programs and providers. Moreover, the disincentives created by high-stakes accountability inevitably lead to goal distortion, wherein achieving the metric becomes the goal irrespective of the processes employed to attain it.

As Stecher and Kirby (2004) illustrate, outcome-based accountability systems frequently induce educators to target the metric or distorted goal rather than the broad public policy goal, resulting in such practices as teaching to the test, cohort manipulation, and disproportionate curricular focus, to name but a few. Such practices not only call into question the efficacy of accountability systems to improve educational practices, but also raise questions related to the validity of improvements reported under such systems.

The basic premise of most state accountability systems as well as the federal system under NCLB is that incentives, in the form of rewards and/or sanctions will "encourage" improved practices that will translate into improved educational outcomes through a cycle of continuous improvement. See Figure 1 below.

Continuous Improvement Cycle (Adapted)

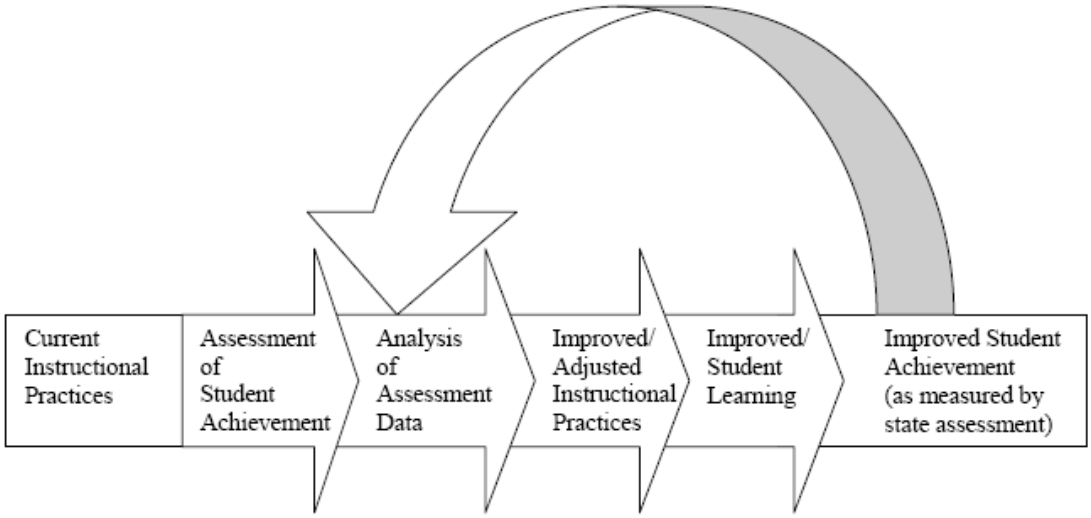

Figure #1

However, as incentive theory has consistently demonstrated in other fields, especially health care and governmental services, actors under accountability pressure tend to distort goals to focus on metrics through practices that do not necessarily, and in fact, rarely support the broader public policy goals — in this case improved educational practices. As Holmstrom and Milgrom (1991) have shown in economic models, the greater the "risks" of failure, such as the challenges in an educational setting, especially input disparity in the form of language, economic, and mobility issues, "the more likely actors are to distort processes" to meet a distorted goal, thereby meeting the goal by attaining an indirect metric rather than the broad policy goal of, for example, improved student learning (p. 34). See Figure 2 below.

Distorted Processes Model

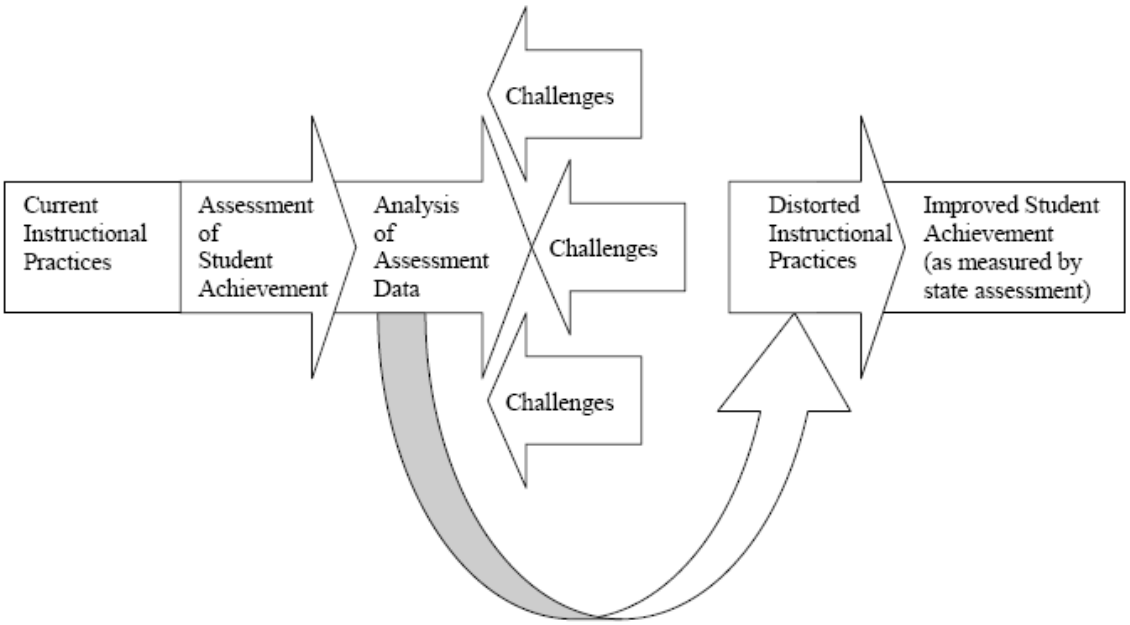

Figure #2

As illustrated in Figure 2 above, substantial risk of failure (denoted therein as challenges) may cause actors to subvert rather than address challenges to meeting a metric (in the case of accountability: student achievement as determined by the rates of students meeting minimum standards on the state assessment). Economists Laffont and Martimort (2001) have noted goal distortion as a problem particularly with accountability systems based on narrowly defined indirect metrics rather than systems which rely on more "holistic evaluative techniques" based on observation, work product, and other factors which may have influenced performance (p. 292). Intuitively, it would seem, such distortions would only be exacerbated in accountability systems such as AEIS and AYP that rely disproportionately on the single indicator of test scores. McNeil (2005) equates such reliance on single indicators to the well-publicized "accounting debacles" that took place related to the collapse of one-time oil giant Enron (p. 65). She points out that Enron utilized a single metric indicator (the company's stock price) to market the

company's success. The company's balance sheets reflected high revenues and profits, thus driving the price up, while questionable, and as it turns out illegal, accounting practices hid the company's massive debt obligations, thereby artificially inflating the company's net worth. Similarly, McNeil (2005) criticizes the heavy reliance of the Texas accountability on the state's standardized assessment, noting it has led to a bevy of questionable practices such as test drill, restricted curricula, and targeted exemptions of students.

The Texas accountability system, as Rothstein (2008) points out, makes little adjustment for so-called risk factors. Though it does disaggregate results by subgroup, overall score requirements are not adjusted for high numbers of students within risk groups. Therefore, schools and districts, such as those along the Texas-Mexican border, with high populations of students in focus groups, including limited English proficient, migrant, and economically disadvantaged, are disproportionately impacted since these populations constitute a great majority of their students. Many schools in the Far West Texas ISD, for example, have economically disadvantaged populations approaching 100 percent, while the district as a whole most recently reported more than 68 percent of its students as economically disadvantaged and nearly a third as limited English proficient (Texas Education Agency, Pocket Edition of School Statistics, FWTISD Supplement, 2007-2008). The federal system is even more onerous, not only holding schools and school districts to a single AYP criteria for overall performance regardless of the population of students in focus groups, but also requiring such standard individually for each and every subgroup category above a minimum population threshold (10 percent of student population with minimum of 50 students).

These statistics bolster the argument that both the state and federal accountability systems are inequitable in their treatment of schools and school districts which are demonstrably

disproportionately challenged to meet accountability requirements under both models (Kim & Sunderman, 2005; Tracey, Sunderman, & Orfield, 2005). Challenges of this type widely impact these schools and districts in terms of personnel, program budgets, and curriculum focus. However, many proponents of test-based accountability continue to insist that the system acts as an incentive to encourage students, parents, teachers, and administrators to work harder to identify and remediate struggling students (Whitehurst, 2010). Such advocates believe student achievement can be improved by raising student motivation, increasing parental involvement, and encouraging states and districts to improve curriculum and pedagogy. This study employed quantitative methods to examine student achievement in schools with varying degrees of accountability risk as a means to determine whether the student achievement gains celebrated by proponents mask distorted instructional practices that have had a disproportionate impact on the most needful students and campuses.

Purpose of the Study

The purpose of this study was to determine (1) whether student performance as measured on the state assessment evidenced significant differences in performance of students at schools related to the schools' level of accountability risk as defined herein, and (2) whether such differences, if any, reflected the prevailing wisdom related to distortion of educational goals and practices, specifically the nature of the curriculum that educational practitioners choose to focus instruction upon as well as the instructional delivery focus related to the target student group. In other words, as is a common complaint related to accountability: Has the use of standardized testing led to narrowing of the curriculum and a focus on the most easily attainable skills delivered at the minimal level to achieve proficiency to thereby achieve as high a possible rate of students meeting at least minimum proficiency levels, as such levels are typically the focus of

high stakes accountability systems, including both the AEIS system and the AYP provisions of NCLB?

Significance of the Study

With NCLB legislation currently stalled as to reauthorization, and as Congress debates the relative positives and negatives of this legislation, this study is significant in that it sought to determine whether quantitative evidence from the accountability instrument's own chief metric could support the theory often alleged anecdotally and qualitatively that NCLB and similar accountability instruments have induced educators to distort educational goals and practices in response to such accountability risks. In addition, the study aimed to determine the extent to which, if such distortion is indeed prevalent, such use of singular assessment metrics for accountability are reliable, let alone appropriate measures by which to determine school efficacy, student achievement, and teacher quality. Though studies by Campbell (1979) and Simon (1978), discussed in more detail later in this dissertation, suggest that public accountability systems are difficult to design due to the frequent misalignment of goals and measurement metrics, it seems unlikely that outcome-based accountability has run its course. Therefore, to the extent that identifying and understanding how accountability systems relate to broad public policy goals and how they manifest unforeseen and unintended consequences, may, it is hoped, mitigate such effects in future systems.

Finally, this study is significant because if, as the model herein suggests, the accountability system intended to incentivize educators to improve instructional practice is in fact, acting to inhibit sound instructional practice in an effort to achieve metrics only loosely aligned to broad policy goals, then in a practical sense, the data (i.e., test scores) used to evaluate curriculum and instructional practice, as well as to formulate much educational policy, does not

in fact accurately reflect either student performance or that of the professionals charged with educating them.

Limitations of the Study

The extent to which findings from this study can be generalized across other student populations may be limited due to the relative small sample of school performance data analyzed in the study. In addition, while the methodological approach employed within this study attempts to mitigate the possibility that differences in student performance observed in relation to the different risk groups analyzed herein may potentially relate to other factors, including the lack of homogenous educational practices, approaches, and backgrounds, such factors cannot be entirely discounted. However, to the extent that student achievement is a function of untold numbers of factors, not the least of which are nonpersistent factors described later in this dissertation, the researcher believes the strong relationship between predicted outcomes and accountability risks make it likely that high degrees of accountability risk are in fact, contributing to the performance measured herein.

In addition, this study relies on state assessment data, the reliability of which is called into question within this dissertation, itself. However, as is evidenced in the discussion dealing with summative and formative assessment, the faults of the data lie principally in their volatility and verifiability. To this extent, although such data is called into question as a metric for reliability purposes, its use as a general indicator of student response, especially in large populations is less volatile and problematic because the study does not seek to determine the educational performance of any single student, teacher, or even campus based on such snapshot data, but rather the performance in a broader sense of campus groups on test items relative to

their position on a continuum of difficulty determined by statewide performance of students on the individual items.

Finally, it is both ironic and apropos that this study relies on test data to make conclusions regarding instructional practice. As this study sought to quantify the instructional impact of accountability through a comparison of the state's own data, it therefore did not rely on any direct measure of instructional practice in the classrooms of the schools from which the data originated. Though the reliability of conclusions drawn from such indirect metrics is a salient point in the theoretical framework upon which the study is based, this only further serves to illustrate the varying conclusions that can be drawn from data relevant to the disparate lenses under which it is examined. Therefore, while the conclusions drawn from the data in this study may be subject to similar discrepancies, the underlying point that metrics of this nature cannot be used to reliably evaluate school efficacy, is, nevertheless, strengthened.

Chapter Summary

The federal school accountability system under NCLB is an overwhelming mandate that impacts hundreds of millions of students across the United States. The legislation, as its title implies, was designed to encourage educators to focus instructional practice such that traditionally overlooked groups of students, particularly students who are economically disadvantaged or who are limited English proficient. However, consistent with Campbell (1979) and Simon's (1978) theories of performance measurement, and Laffont and Martimort's (2001) studies involving incentives, many researchers such as Linn (2009) and Rothstein (2008) now question the premise that accountability has improved or will improve instructional practice in schools, pointing to conflicting statistics on other measurement instruments, for example, NAEP scores.

This chapter provided background information, a statement of the problem, the theoretical framework that guided the study, the purpose and significance of the study, the research questions undertaken by the study, and a definition of key terms used in the study, as well as the study's limitations. In Chapter 2, the researcher will review the literature related to assessment and accountability, including the legal framework and mandate for principal leaders to utilize assessment and data to inform instructional practice in schools, the role of accountability systems as an incentive to improve instructional practice in the public schools, and the concept of goal distortion as it applies to organizations in general and to schools in particular, including such commonly criticized practices as narrowing of the curriculum, teaching to the test, and the impact of test-focused practices on high-risk groups, including economically disadvantaged students and those who are limited English proficient.

Chapter 2

REVIEW OF LITERATURE

Introduction

The review of literature is focused on the following areas: the legal framework underpinning the role of assessment and the use of data to inform instructional practice, the perceived impact of accountability on instructional practice, the concept of goal distortion in general and as applied specifically to education, implications for special populations, and finally requirements under the accountability system.

Legal Framework

In Texas, standards for educator certification for school administrators are contained within the Texas Administrative Code, Title 19, Part 7, Chapter 241, Rule §241.15. The code represents the legislative framework that delineates the duties and responsibilities of administrators related to the instruction of public school students in Texas. The expectation that school leaders in Texas promote the analysis of student data to adapt and guide instruction is set forth therein as follows:

> I Learner-Centered Leadership and Campus Culture. A principal is an educational leader who promotes the success of all students and shapes campus culture by facilitating the development, articulation, implementation, and stewardship of a vision of learning that is shared and supported by the school community. At the campus level, a principal understands, values, and is able to:
>
> (5) utilize emerging issues, trends, demographic data, knowledge of systems, campus climate inventories, *student learning data*, and other information to develop a campus vision and plan to implement the vision.
>
> (g) Learner-Centered Curriculum Planning and Development. A principal is an educational leader who promotes the success of all students by facilitating the design and implementation of curricula and strategic plans that enhance teaching and learning; alignment of curriculum, curriculum resources, and assessment; and the use

of various forms of assessment to measure student performance. At the campus level, a principal understands, values, and is able to:

(1) use emerging issues, occupational and economic trends, demographic data, *student learning data*, motivation theory, learning theory, legal requirements, and other information as a basis for campus curriculum planning.

(h) Learner-Centered Instructional Leadership and Management. A principal is an educational leader who promotes the success of all students by advocating, nurturing, and sustaining a campus culture and instructional program conducive to student learning and staff professional growth. At the campus level, a principal understands, values, and is able to:

(4) utilize interpretation of *formative and summative data* from a comprehensive student assessment program to develop, support, and improve campus instructional strategies and goals.

The establishment in schools of processes to analyze student learning data for the purpose of adjusting instruction is also recognized in school leadership standards set forth by the Interstate School Leaders Licensure Consortium as follows:

Standard 2

A school administrator is an educational leader who promotes the success of all students by advocating, nurturing, and sustaining a school culture and instructional program conducive to student learning and staff professional growth.

Knowledge

The administrator has knowledge and understanding of *(in part)*:

- *measurement, evaluation, and assessment strategies*

As illustrated by the codification of these duties and responsibilities above, the utilization of state assessment data to inform and adapt instruction is clearly supported by a legal framework. However, the duties and responsibilities below clearly mandate that administrators

provide strategies and structures to meet broad educational goals and that such strategies and structures be developed based on a variety of information sources to meet individual student learning needs. In Texas Administrative Code, Title 19, Part 7, Chapter 241, Rule §241.15:

> (g) Learner-Centered Curriculum Planning and Development. A principal is an educational leader who promotes the success of all students by facilitating the design and implementation of curricula and strategic plans that enhance teaching and learning; alignment of curriculum, curriculum resources, and assessment; and the use of various forms of assessment to measure student performance. At the campus level, a principal understands, values, and is able to:
>
> (3) implement special campus programs to ensure that all students are provided quality, flexible instructional programs and services to meet individual student needs.

And in the Interstate School Leaders Licensure Consortium:

> Standard 2
>
> Dispositions
>
> The administrator believes in, values, and is committed to *(in part)*:
>
> - student learning as the fundamental purpose of schooling
> - the proposition that all students can learn
> - professional development as an integral part of school improvement
>
> Performances
>
> The administrator facilitates processes and engages in activities ensuring that *(in part)*:
>
> - professional development promotes a focus on student learning consistent with the school vision and goals
> - barriers to student learning are identified, clarified, and addressed
> - multiple opportunities to learn are available to all students
> - curriculum decisions are based on research, expertise of teachers, and the recommendations of learned societies
> - a variety of sources of information is used to make decisions

The Impact of Accountability on Instructional Practice

This review of literature will first further examine the concept of goal distortion as a consequence of standardized test-based high-stakes accountability systems, detailing widely criticized practices aimed at targeting the test or gerrymandering student samples, and then analyze key risk factors which disproportionately challenge schools and school districts, notably socioeconomic factors as well as high populations of English language learners.

Goal Distortion and the Law of Performance Measurement

Social science researchers have long observed that workers behave differently when monitored for efficiency. As Gillespie (1991) relates, early industrial studies conducted in the early 1920s at the General Electric factory first uncovered this previously unknown phenomenon. As part of a scientific management approach, researchers Elton Mayo and Fritz Roethlisberger, through what later became known as the Hawthorne Investigation at Western Electric Company, sought to determine the optimal illumination at General Electric plants for workers to be the most productive. They were surprised to observe that workers at the Hawthorne plant increased production both at dimmer and brighter levels of illumination. After conducting interviews with the workers, Mayo and Roethlisberger determined that workers had adjusted their performance as a consequence of the evaluation.

In the case of the Hawthorne experiments, monitoring had a positive unintended consequence. However, the Hawthorne workers, as Rothstein (2008) notes had no personal stake in the results of the study. No incentives were offered for increased production and no sanctions were threatened for decreased production, though one might assume that some workers feared the study results might be used to criticize their job performance. For more overt performance measurement models, however, such assumptions are not necessary. Workers in many

occupations, both public and private, are acutely aware of performance monitoring systems as well as the incentives and disincentives related to measurement outcomes.

A half century later, Northwestern University social scientist Donald Campbell (1979) determined that accountability and control systems "which included elements of possible rewards or punishments created incentives for workers to utilize deception and fraud to appear more competent than they actually were" (p. 43). Campbell articulated his findings in a theory which came to be called the law of performance measurement:

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor. (p. 43)

Campbell, as well as Carnegie-Mellon University professor Herbert A. Simon (1978), defined two fissures which have weakened the foundation upon which public accountability policy has been constructed: the failed attempt to measure complex public goals with a simple quantifiable metric; and the corruption and fraud which have been engendered by attempts to meet such goal.

Even as politicians and policy makers debate the relative merits and deficiencies of NCLB as part of the effort to reauthorize the act, critics assail the attempt with countless examples of such fraud and corruption, running the gamut from the widespread perception that teachers "teach to the test" to more isolated incidents, or so it is thought, of blatant cheating and outright fraud (Jones, Jones, & Hargrove, 2003, Ryan, 2004). Nevertheless, with continued strong public support for some type of accountability system for education, lawmakers are resolutely, albeit possibly futilely, engaged in efforts to draft changes to the law that will address its weaknesses (Education Trust, 2011). In doing so, Rothstein (2008) notes three key obstacles that must be overcome:

(1) Oversimplified outcome measures (i.e., the disproportionate reliance on a single indicator [test score] to determine progress, resulting in goal distortion and manifest process distortion including excessive test prep activities, narrowed curriculum, questionable exemptions, and marginalizing of students at the high and low end; and

(2) Insufficient or absent inputs adjustment structures to account for disparate populations among schools and incumbent disparate impact of the accountability provisions (e.g., to provide a more equitable accountability comparison for schools with high levels of student populations at risk of failure as opposed to those with low populations of such students); and

(3) Untrustworthy statistics derived from assessments subject to probable sampling error associated with teaching to the test as well as the relatively small student cohorts for subgroups and subsequently large confidence intervals in score reporting that further undermine the system's credibility. (p. 6)

Educators have been harshly criticized for seeming to put personal professional goals before student learning, but as Rothstein demonstrates, goal distortion is a phenomenon readily observable in any system, public or private, where incentives and/or disincentives are employed to effect outcomes defined by a simple, indirect metric. In some respect, nevertheless, such criticisms have merit in so much as they involve educators readily adopting practices which they themselves denigrate in pursuit of higher test scores and satisfactory accountability ratings. However, such incongruence, if in fact such practices are being employed on a large scale, only underscores the strong influence that accountability incentives may be exerting on instructional practitioners.

Such pressures, critics argue, are especially strong in schools and school districts which face disparate levels of challenge. Rippberger and Staudt (2003) note, for example, that while

the numbers of students in the United States for whom English is not the first language is dramatically increasing in the nation at large, border states, and especially border communities, face a far greater challenge than schools, districts, and even states which may have student subgroup populations "often too small to be subject to accountability measures" (p. 126). They go on to detail how in schools confronted with such disparate challenges "preparation for tests tends to subvert experience-based learning" and observed the propensity for pencil-paper test drills over participatory learning (p. 129).

Prevalence of Goal Distortion in the Other Sectors

Goal distortion is not unique to educational accountability systems. Long before the passage of NCLB, Holmstrom and Milgrom (1994) first identified the impact of incentives on parties to contracts in the economic sector. As Rothstein (2008) fully details, such distortion has been a predictable occurrence whenever incentives are insufficiently related to determinate metrics and has been widely observed in medical and legal practice as well as governmental services. Such distortion as well as often mismeasurement of outputs and failure to adjust risk according to inputs has encouraged practitioners to adversely redirect effort. For example, Rothstein details how an accountability model designed to assess physician performance related to heart surgery patients led to poorer cardiac care for those patients most in need. Designed to inform patients of the best surgery centers and to encourage improvements in practice, the system which used a single, indirect metric (mortality rates) encouraged doctors to treat the riskiest patients (also presumably the most needy) with conservative approaches such as drug therapy that delayed what often turned out to be necessary surgery.

In another example, Rothstein describes the evaluation of government unemployment counselors under a system that very heavily weighed the number of successful placements in

jobs. Again, goal distortion resulted as meeting the metric became the primary goal of counselors who adopted processes that ironically tended to exclude the most needy applicants. Faced with sharp pressure to maximize the placement of referrals, counselors focused their efforts on highly skilled workers (often the most recent, short-term unemployed) and avoided referring low skilled, long-term unemployed applicants who were less likely to be hired, though the applicants who had been unemployed for longer times were clearly in the most need. This practice is an example of what economists refer to as cream skimming and Rothstein relates it to the targeting of so called bubble students who are the easiest to move into the proficient realm, at the expense of lowest performing students who are deemed to have little or no chance of passing a test but who are clearly in the most need.

In a Rand study, Stecher and Kirby (2004) note additional examples of goal distortion in private enterprise and add an additional caution, explaining that many of the failed accountability models relied on insufficient numbers of indicators which were, nevertheless, "far more numerous than the indicators used for most educational accountability systems" (p. 12).

As educators throughout the nation continue to struggle with accountability under NCLB and state systems, it is important that the laws that focus on risk subgroups be reformed rather than eliminated. For too long, long before accountability systems created the distortions described herein, difficult to educate children, indeed the most needy of our students, have been passively discriminated against through systems that overlooked poor performance in preference of an efficient, scientific management approach, which shuffled students through the system with little regard to their educational progress.

Impact on Schools and Districts with High Defined Subgroup Populations

As noted earlier in this proposal, Texas's accountability system, the Academic Excellence Indicator System (AEIS), has long held schools accountable for performance of their English language learners. However standards for which students would be tested for accountability purposes have vacillated over the years with exemption periods ranging from one to three years before students must be tested for accountability. Additionally, until recently, AEIS, though desegregating data by subgroup (including Limited English Proficient [LEP]), did not set performance standards for individual subgroups as did the AYP component of NCLB, instead relying on these students' presence in the "all students" group as an indicator of academic proficiency. Though at first glance, such policy appears favorable to schools (especially those struggling to meet federal AYP requirements for each subgroup), its structure fails to adjust for the high risk levels of schools with high populations of English language learners. Current AEIS provisions now require separate evaluation of subgroups with a student population of at least 30 students, provided the student group accounts for at least ten percent of the all student population and of subgroups of 50 or more regardless of the percentage of the overall student group (2011 Accountability Manual, p. 53). While these adjustments have made accountability exposure somewhat more equitable, they still do not adjust for the disproportionate exposure to which schools with high populations of students in defined subgroups are subjected.

Title III of NCLB, in contrast, holds states accountable for meeting the same accountability requirements for defined subgroups, including LEP students, which meet specific criteria as measured by state summative assessments based on specific academic content standards as well as for progress toward English proficiency. Though at its most drastic stages,

the accountability system can impose harsh sanctions, including loss of federal funding, for failure to meet established proficiency standards, schools in the early stages of AYP sanctions in fact receive additional funding to assist with the remediation of students, though the funds come with a bevy of strings attached. However, continued failure to achieve standards can result in the loss of federal funding and reassignment of school personnel, among other sanctions (NCLB Title III, Subpart 2, Sec. 3122).

While neither state nor federal legislation dictates the adoption of specific educational programs for students belonging to defined focus groups, NCLB does require that the educational programs that districts select be designed upon sound, research-based practices and established educational theory, provide trained personnel with appropriate instructional materials, and include a reliable evaluation process (Lessow-Hurley, 2003).

For example, research has shown that LEP students achieve higher levels of academic success when quality instruction is provided in both English and their native language (Bailey & Butler, 2003). Learning is further facilitated by instruction that incorporates elements of both the students' native culture as well as learning about the dominant culture and others. Conversely, students immersed in an unknown language and culture struggle academically when schools pressure them to abandon their native language and culture (Cummins, 1995).

According to the California State Department of Education, LEP students programs should observe the following framework:

- instruction and support in the student's primary language
- instruction for English language development
- sheltered academic content, and
- multicultural representative curriculum

Assessment of learning should measure LEP students' growth in each aspect of the framework. Assessments should be free of bias and be valid and reliable. Results should be analyzed and used to adjust instruction to best support students' progress toward their goals (Freeman & Freeman, 1998).

Hence, studies indicate that effective bilingual education programs promote high academic achievement for LEP students (Thomas & Collier, 1997). With a hot debate currently raging across the nation related to immigration, many citizens in the United States oppose bilingual instructional programs (Reyhner & Singh, 2010). For example, voters in California, Arizona and Massachusetts, have recently passed referendums that severely limit, if not effectively eliminate, most bilingual programs in favor of structured English immersion where LEP students receive all instruction in English. Though key provisions of these laws are under court review, the mood of the country reflects a growing dissatisfaction with bilingual programs. Even the names of national agencies charged to oversee instructional programs in this area are being revised to reflect the changing political climate. For example, the Office of Bilingual Education and Minority Language Affairs (OBEMLA) is now called the Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students (OELA). Likewise, the National Clearinghouse for Bilingual Education is now the National Clearinghouse for English Language Acquisition & Language Instruction Educational Programs.

Although NCLB does not prohibit native language instruction for immigrant students, the current focus of education for LEP students clearly emphasizes English language teaching over providing access to content or literacy in students' primary language. Unfortunately, while NCLB permits testing in students' primary language, most states do not provide for such testing

(or, as in the case of Texas, provide limited opportunities for native language testing) (Bratt & Sunderman, 2005).

Moreover, development of reliable and valid assessments (in English) for LEP populations poses special problems for test developers who must accurately document LEP students' academic progress despite their limited English proficiency, cultural disconnects and lack of opportunity to learn the material being tested. Too often, academic assessments of LEP students are, practically speaking, measures of students' English language abilities measured by their comprehension of the questions, rather than their purported level of knowledge of the content area assessed. Thus, as noted throughout this proposal, decisions made by teachers and administrators based on the state scores of LEP students' on standardized tests, run the risk of distorting educational processes as well as the garnering inappropriate, inequitable judgments.

In a 2003 position paper related to the testing of LEP students, the professional organization Teachers of English to Speakers of Other Languages (TESOL) held that:

> [i]nasmuch as [standardized] tests measure content in combination with linguistic abilities, English language learners are at a distinct disadvantage that is difficult to accommodate. Further, cultural differences and limitations concerning opportunity to learn can lead to unfair interpretations of low test scores and assessment discrimination…[S]ince high English proficiency is a prerequisite for success on high-stakes tests, such assessments are not appropriate for English language learners and often do more harm than good. (p. 3)

Likewise other professional organizations, including the American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education also caution against decision-making based on the scores of English Language Learners on standardized tests:

> …test norms based on native speakers of English either should not be used with individuals whose first language is not English or such individuals' test results should be interpreted as reflecting in part current level of English proficiency rather than ability, potential, aptitude or personality characteristics or symptomatology. (*Standards for Educational and Psychological Testing, 1999,* p. 91)

On the other hand, proponents of standardized testing (Whitehurst, 2010) maintain that assessments can inform individual students as well as their teachers and parents regarding their academic achievement relative to other students as well as identify groups of students who are struggling so as to help schools and districts understand the fundamental strengths and weaknesses of their students. However, even strong proponents of standardized tests acknowledge the limitations of assessment scores and caution that such scores should be only one factor among many that educators use to inform instruction (Farr & Trumbull, 1997).

Thus, the use of a single test score to determine a student's level of proficiency, educational placement, or other important consequence, is considered 'high-stakes.' The high-stakes nature of accountability under NCLB has been found to strongly effect the educational environment in U.S. schools (Wright, 2002). As described earlier, American schools are rewarded or sanctioned based on the performance of their students on such assessments in accordance to AYP proficiency requirements. Researchers have documented that low-performing schools have narrowed their curriculum, and limited, excluded, or de-emphasized subjects that will not be tested (McNeil, 2000). As schools struggle to demonstrate proficiency, students in low-performing schools are subjected to hours of monotonous test prep.

Ironically, for legislation designed at least in part to address educational gaps between White students and minority and economically disadvantaged students, the impact of the high-stakes accountability measures pursuant to NCLB has been particularly troubling for students in

these socioeconomic focus groups, including LEP students, migrant students, and economically disadvantaged students. Schools with high populations of these groups are disproportionately exposed and not surprisingly then disproportionally represented in the low-performing categories of both the state and federal accountability systems (Routledge, 2003).

Given that subgroup performance is a vital cog in the NCLB legislation and the plank on which many educational entities are most challenged when it comes to meeting AYP standards, it is useful to examine the makeup of the accountability system, particularly as it pertains to subgroup representation and expectation, and also to look at some of the troubling aspects of the instrument itself that relate to its reliability as a measure of educational achievement.

Measuring Up Under NCLB Federal AYP Provisions and Guidelines

As the premise of this study postulates that the high-stakes accountability provisions of NCLB and other stringent accountability systems, including especially Texas's AEIS as pertains to this study, have distorted instructional practice in American public schools, it is useful to carefully examine the provisions of the accountability instrument so as to better gauge the accountability pressures that schools are subjected to and which may result in said instructional distortion, especially with reference to subgroup provisions and the determination of proficiency both for focus subgroup and all students groups.

Under NCLB four types of student subgroups are defined which include students from racial/ethnic groups (American Indian, Asian, Hispanic, African American, and White), students identified as Limited English Proficient, students with disabilities, and students identified as Economically Disadvantaged (typically determined by eligibility for the Federal Free and Reduced Lunch Program). NCLB mandates that schools meet AMOs (Annual Measurement Objectives) in all subgroups where there is a numerically significant (i.e., statistically reliable

sample population size) in mathematics and reading and that 95 percent of all students, as well as 95 percent of each qualifying subgroup, participate in testing in order for a school or school district to meet federal AYP requirements (2011 Adequate Yearly Progress Guide).

Under NCLB (20 USC 6311 (b)(2)(C)(V)(II), however, AYP is not required when student populations in a defined subgroup are "insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student." However, NCLB gives states broad leeway to determine what qualifies as a statistically reliable student subgroup (Erpenbach, Forte-Fast, & Potts, 2003). Thus states have broad discretion to determine the minimum number of students that a school must have in a subgroup to trigger NCLB subgroup reporting and tracking provisions as well as AYP subgroup performance requirements. Current NCLB subgroup requirements for AYP purposes range from 5 (Maryland) to 60 (Kentucky), with the mode being 40. In Texas, for example a subgroup must be at least 50 members and comprise at least 10 percent of all students or at least 200 members regardless of percentage to be measured for AYP purposes (compared to 30 members/10 percent or 50 members for the state system) while in California schools must have at least 50 students in a subgroup and comprise at least 15 percent of all students with subgroups of 100 or more evaluated regardless of their percentage of all students. Groups not meeting minimum size requirements are not counted separately for accountability purposes.

Thus, provisions of accountability systems exclude to various degrees students in focus groups. For example, in 2006, nearly two million students from defined subgroups were excluded from AYP calculation under the various state standards for minimum size. Likewise, while proficiency standards under NCLB continue to rise (the "moving targets"), consistent with the 100 percent proficiency goal of the program, nearly half the states (23) have increased the

minimum size requirements to include subgroups for AYP calculation since 2004, while no states had decreased the size (Chudowsky & Chudowsky, 2005).

The clear implication is, of course, that these changes to state accountability provisions are made to facilitate more schools successfully meeting AYP requirements, a theory supported by several studies which have shown that, in general, the percentage of a state's schools that meet AYP increases as the minimum subgroup size increases (Porter, Linn & Trimble, 2005; Simpson, Gong & Marion, 2006).

Separate and Unequal: Moving Targets and an Unattainable Goal

As noted earlier, student achievement, or lack thereof, as pertains to the AYP provisions of NCLB and other high-stakes accountability instruments, is measured by the percentage of students meeting proficiency standards on the state's summative assessment and whether or not the percentage meets the instrument's Annual Measureable Objectives ("AMOs"), which specify the minimum percentages of students who must perform at or above the state's proficiency level or "cut score" for both reading and mathematics (NCLB, PL-107-110, Title II, Part A, Subpart 4, section 2141)..

NCLB critics, including Clarke (2007) and Cohen and Moffitt (2009) have long pointed to validity concerns with AYP accountability arising from the fact that the instrument relies on the individual states' independently developed and widely variable state assessments, developed from equally disparate and inconsistent state standards. Further complicating the potential for discrepancy, states enjoy autonomy for determining what level of performance constitutes proficiency, meaning states are allowed to set their own cut points for passing. This results in assessments of different rigor and substance being evaluated differently from state to state. Based on a comparison of state standards and the performance of states on the NAEP test,

researchers Peterson and Hess (2006) concluded that what constitutes a proficient student varies widely from state to state. Based on this analysis they concluded that a primary reason some states exhibit high proficiency rates relates to the relative weakness of the state's proficiency standards and vise-versa.

While it may be tempting to dismiss talk of standards discrepancy between states as mere academic rivalry, the differences between what constitutes proficient in one state compared to another are far from trivial. In a 2003 study, researchers at the National Center for Education Statistics compared the state proficiency levels of campuses with the NAEP performance for the same campuses as a means of correlating the state standards to the NAEP. McLaughlin et al. (2008) determined that equivalent NAEP performance varied significantly for states' proficient level students, estimating that in some cases as much as 20 percent of one state's proficient students would not have scored at the proficient level on another state's summative assessment.

Likewise, differences in state AMOs are often far greater than most would assume, despite the fact that all states, theoretically, are subject to the same ambitious, and many would argue unrealistic, broad NCLB goal of 100 percent proficiency by 2014. Under NCLB provisions, states must establish yearly AMOs building toward the 100 percent goal. Though NCLB does mandate that intermediate goals remain fixed for no longer than three years, states may choose how often to adjust the scores (annually, biannually, or every three years), resulting in a mish mash of growth trajectories and requirements, further adding to the inequity of the system for schools in different states. Porter, Linn and Trimble (2005) identified four basic trajectories types pertaining to intermediate AMOs: (1) straight line, with consistent annual increase; (2) stair-step with consistent increases every three years; (3) front-loaded, with large increases in early years, and (4) back-loaded, with the larger increases delayed until later years.

While some states set the intermediate AMOs, often referred to as the "Moving Targets," at equal intervals of growth (e.g., Florida), nearly half of all states elected to back-load larger score gain requirements to later years, probably in hope that the goal would be moderated in the subsequent reauthorization of NCLB.

As 2014 rapidly approaches, and with projections that more than 90 percent of the schools in some states will miss AYP over the next three years (Wiley, Mathis, and Garcia, 2005), the back-loaded approach may prove itself prudent. In fact, President Barak Obama has now decided to issue by executive order, waivers that relieve schools of these AYP provisions under certain guidelines. However, it is unclear how the waivers will work with respect to the law as congress continues to work on its reauthorization. Nevertheless, currently, most states/districts are still required to meet established AMOs. Just how disparate the impact of the AMO provisions of NCLB are, even discounting previously described discrepancies in the state assessments themselves, their passing thresholds, and the standards upon which they are based, is evident through an examination of the initial AMOs set by states in 2002. The initial AMOs (the starting point for achievement growth required under AYP) were established based on the larger of the percentage of students at the proficient level in a state's lowest achieving subgroup or in schools at the 20th percentile in the state (calculated based on enrollment, with schools ranked based on the percentage of students rated proficient or above). Such criterion resulted in some states setting initial AYP AMOs as low as 7 percent compared to 75 percent on the same subject/grade level in other states. In other words schools in one state where as few as 7 percent of students achieved proficiency on a state assessment might be determined to meet AYP for that subject/grade level, while schools in another state where up to 74 percent of students achieved proficient ratings might be deemed to have "missed AYP."

In its current state, NCLB's accountability ratings bear little if any resemblance to what is actually taking place in America's schools. An instrument that relies on inconsistent inputs, interpreted and applied inconsistently, against inconsistent benchmarks cannot pretend to reliably compare the educational efficacy of schools. Add on top of these problems an end goal that most observers agree is unattainable, though laudable. As Rebell and Wolff (2008) have noted, the 100 percent proficiency rate is far from mere motivational diatribe, being the legal mandate that underpins NCLB's accountability structure—a mandate the authors contend has "never been achieved in history" and the feasibility of which "has never been demonstrated" (p. 5). The authors cite Linn's (2004) findings that in order to achieve such an ambitious target, students in some subject/grade level (e.g., 4th grade math) would have to progress at a rate nearly 16 times the national average over the period 1998 to 2003. Though lauding NCLB's attempt at ensuring that the educational needs of all students are addressed, the authors add that the unempirical goal of 100 percent proficiency is not only irrational, but in the sense that it will soon subject thousands of schools around the country to a "failing" rating and the problems that accompany such identification, it is "causing considerable harm" (p. 6)

Safe Harbor, Confidence Intervals, and Projected Growth

Besides allowing the states broad authority to set standards, design assessments, set cut points and proficiency ratings, many states continue to struggle to meet AYP requirements, especially as the Moving Targets increase over time, since NCLB designates a series of increasingly-severe sanctions for schools that miss AYP. Such sanctions take affect based on the number of consecutive years schools miss AYP, however, the legislation seeks to soften the landing, so to speak, by allowing a variety of exceptions to the base-line AMO requirements. The most well-known of these exceptions is commonly referred to as "safe harbor," which

allows for schools that miss AYP to be deemed to meet AYP guidelines provided they show sufficient progress toward meeting the AMO goals within ten years, projecting the same growth rate, provided certain other conditions for graduation and/or attendance rates are also met.

States may also choose to use confidence intervals (sometimes known as "margin of error") in order to determine whether schools meet an AMO. This provision allows a school that scores below a certain AMO, whether at the "all students" or a subgroup level, to be deemed to meet AYP provided its performance falls within the confidence interval surrounding a specific proficiency goal. States may even use confidence intervals to determine which schools meet AYP's safe harbor provision, thus allowing an exception to meet an exception. As the pressure mounts for states to meet their own Moving Targets, more and more of them have adopted the use of confidence intervals to calculate AYP. By 2003, more than half of all states (31) had already adopted the use of confidence intervals. By 2009, that number had ballooned to 45 states.

A confidence interval is basically a statistical measure of how a result might vary were a given population to be re-sampled. Confidence intervals basically attempt to account for sampling error, such as mitigating the effects of distracted students, which may occur during testing. In this case, the greater the confidence required, the greater the confidence interval will be. States, such as Arizona, which demand a greater confidence level (99%) are actually more lenient because this creates a larger confidence interval necessary for achieving such confidence than the typical 95 percent confidence level. However, even assuming a 95 percent confidence level is calculated for a given school's (Sample Campus) proficiency rate (e.g., 65%) resulting in a confidence interval of plus or minus 6 points, means that Sample Campus could meet an AMO of up to 71 percent (the school's actual performance, plus or minus 6 points). As Rogosa (2003) has argued, however, the very concept of confidence intervals is counterintuitive to a

proficiency requirement, because in the case of the Sample Campus described above, the school's so-called "real" proficiency rate is equally likely to be six points less its actual score (59%) as it is to be six points greater (71%).

However, for purposes of AYP sampling error is always calculated to benefit the target campus. Further as Cronin, et al (2009) have contended, the notion of sampling error is incongruent with the realities of testing under NCLB, pointing out that sampling error as applied to, for example, opinion polls, is normally employed to account for the relatively small sample of an overall population that is actually sampled and from which assumptions are based. However, under NCLB schools must test at least 95 percent of their students, meaning that the overall population, or nearly so, is in fact sampled, clearly outside the scope of the most common justification of the use of confidence intervals. Supporters of such intervals argue instead that the confidence intervals are necessary because the sampling represents a snapshot of student performance at an isolated time. However, as the researchers argue, "under such a broad application, no number could ever be seen as determinative for any reason" (Cronin, 2009, p. 16). Imagine, they ask, if election results were required to meet confidence level parameters because the election is a sample of public opinion on an isolated date.

In addition to safe harbor and confidence intervals, fifteen states either utilize or have utilized a growth projection model (Alaska, Arizona, Arkansas, Colorado, Delaware, Florida, Iowa, Michigan, Minnesota, Missouri, North Carolina, Ohio, Pennsylvania, Tennessee, and Texas) which allows the states to count as proficient students who do not currently meet proficient levels based on their assessment results but who are projected to meet proficiency levels at a future testing date, based on a complex model that looks at student growth over time and historical school performance. An analysis by the National Opinion Research Center found

that use of the projection model in these states allowed an average of 9 percent more campuses to meet AYP provisions than would have without the model (Hoffer, et al, 2011). Depending on how the growth models were formulated, there was wide variation in how much the growth models aided states in meeting AYP, with some improving only a small percentage while others (e.g., Ohio) improved their numbers of campuses meeting AYP by more than a third (34 percent). However, the study noted that preliminary reviews as to the accuracy of such projections show that under the best projection models slightly more than 20 percent of students projected proficient at a future date remained below proficient levels at the target date, while the accuracy of most projection models hovered around 50 percent. Texas has recently discontinued use of its growth model, the Texas Projection Measure ("TPM") after state education agency officials acknowledged that the measure incorrectly projected proficiency as much as 50 percent of the time depending on the assessment, grade level, and subject area (Thevenot, 2010). However, the TPM was utilized to determine accountability for several years in Texas and similar measures continue to be utilized in other states. These projection measures are used in conjunction with both safe harbor and confidence interval exceptions, meaning in effect, a school could potentially use a growth model to count students who were not proficient as proficient in order to move within the confidence interval of a safe harbor provision. In essence, this would build a three-tiered ladder of exceptions, although it is clear that many of the students in the growth model will not attain proficiency as projected, and although the confidence intervals are deeply flawed as described earlier in this section.

Notwithstanding these problems, the fact that only fifteen states use growth models means that students and schools in the other 35 states who are being measured under the same instrument may in fact demonstrate vastly different performance levels than their schools'

accountability ratings might suggest. The growth models only add to the inconsistency by which schools and school districts are evaluated under the federal accountability measure of AYP.

Score Volatility and Noisy Data

Prior to NCLB's implementation, critics of accountability systems dependant on single-score results from large scale standardized state assessments were already voicing concerns that metrics collected from such assessments were compromised and thus unreliable for determining school efficacy due to the high degree of score volatility unrelated to persistent (i.e., policy or instructional) factors. In one study of score volatility, Kane and Staiger (2002) of the Brookings Institution, found that, depending on the size of the school, between 70 and 80 percent of score volatility was the result of non-persistent factors (e.g., some type of distraction on test day) rather than persistent factors (e.g., a new teacher, teaching approach, educational program, etc.). The study results suggest, therefore, that gains/losses posted by schools year over year are much more likely to reflect chance factors rather than systemic, duplicable educational reform (Kane & Staiger, 2002).

Kane and Staiger's (2002) study tracked the assessment performance of several hundred North Carolina elementary students over a multi-year period and compared variability in their scores from the standard mean at both the school and state levels. The authors note that a substantial amount of variability can be explained as sampling error. The authors further explain that this is especially prevalent given the typically small cohorts of elementary students at a particular grade at any given school (e.g., North Carolina schools averaged about 65 students per grade level cohort) and argue that much variability potentially exists in the sample simply due to the fact that the students in the sample are different (Kane & Staiger, 2002). The authors project that in schools which are particularly heterogeneous, such variability will likely be much greater.

The authors then compare this to variability at the state level which theoretically should be much smaller due to the much larger sample. However, data indicates that the state-level variation is nearly as large, consistent with the idea that most variation is due to outside factors and that only between 12 and 16 percent of variability is due to empirical differences in schools.

The authors note also that states' efforts to address preexisting factors in the form of value-added measures display similar, and in fact exasperated trends. Again, the variability in score gains for students within a particular school was only slightly larger than the variability overall, indicating again that such gains are mostly explained by outside factors beyond a school's control such as outside distractions or weather. This conclusion is supported by data from North Carolina that ranked the top ten improving schools each year for a decade with only nine (9) out of 101 schools ranked (one year two schools tied for 10th) appearing twice and only one school appearing three times. In other words, schools had trouble repeating performance gains because, following the author's hypothesis, such gains likely resulted from chance rather than actual differences in how the schools delivered instruction. As an anecdotal example, the authors noted a Massachusetts example where a district was lauded for remarkable gains in its 10th grade scores with extensive newspaper coverage of policy and instructional changes the district had implemented. Further study, however, revealed that the district had tested only 26 tenth grade students and comparable variability could be found by randomly sampling any 26 students across the state (Kane & Staiger, 2002).

Linn and Haug (2002) conducted a similar study on Colorado students tracking the percentage of students who achieved the proficient or advanced level year over year. Linn and Haug (2002) found that student performance for a given year/test was strongly linked to that student's performance the previous year regardless of school or teacher. The study revealed that

such trends were true across cohorts as well. For example, the performance of a cohort of students was strongly linked to that group's previous year's performance regardless of teacher/school placement. The authors also noted, likewise, that volatility in change scores from year to year at campuses. That is, the instability of changes from year to year, with schools showing strong gains significantly more likely to show smaller gains or even declines the following year, and school showing declines one year much more likely to post increases the following year, all of which seems to argue against the idea that such gains/losses reflect school instructional practices. The Colorado data shows substantial between school variability as would be expected with regards to raw scores but the within school variability remains consistent and highly correlated to previous years scores. In other words, cohorts of students advance through schools along very predictable performance paths which do not show much variation as a result of educational programs or practices. Similar to the findings of Kane and Staiger (2002), Linn and Haug (2002) conclude that the use of successive cohorts' performance on standardize assessments is an unreliable metric for determining school, district, or teacher efficacy due to the volatility of non-persistent factors (Linn and Haug call this "noise") to which such results are highly susceptible.

Chapter Summary

This chapter reviewed the literature related to the legal framework by which Texas administrators are required to utilize data from state assessments to inform instruction, as a discussion of the impact of federal accountability on instructional practice. The concept of goal distortion was introduced as related to both educational organizations and in other sectors, and examples were covered from other sectors. The researcher also presented literature related to the accountability system's impact on schools with large populations of defined subgroups subject to

the subgroup provisions of AYP. Finally, the chapter provided information about the various disparate requirements under NCLB as well as research that calls into question the validity of the test scores themselves.

In the following chapter, the researcher will discuss the methodology used in the study, including how participants were selected, defined in risk group categories, and tested for differences in item level performance on the 2011 administration of TAKS reading and math. Information about the simple linear regression test performed with SPSS statistical software is also provided.

Chapter 3

METHODOLOGY

This chapter will provide a brief overview of the study, including a review of the purpose of the study, a restatement of the research questions, as well as an explanation of why the researcher chose the investigative model that was utilized. The chapter describes any ethical considerations that were important to the study, as well as a description of how the subjects were chosen and assigned to risk groups. Finally the research design and plan for data analysis is described.

Overview

The purpose of the study was to determine whether evidence could be extracted from the results of state standardized tests which supports the theory that test-based high-stakes accountability models, specifically the federal AYP provisions of NCLB, have negatively impacted or distorted instructional practices by inducing schools to target curriculum and focus instruction at levels designed to realize the maximum increase in students achieving minimum proficiency, without regard to such practice's impact on students other than those in the focus/targeted group level. The specific research questions this study proposed to answer are:

1. Does state assessment data support the theory that high stakes accountability systems encourage educators to disproportionately direct (distort) instructional practices to minimum skills levels?
2 Does the distortion of instructional practices, if any, increase subject to the accountability exposure of schools?

This section describes and justifies the research methodology and research methods proposed to answer the above noted research questions.

Research methodology is a generic term that refers to the general logic and theoretical perspective of the research project (Bogdan & Biklen, 2003). Methodology is used to describe the theory of how the research should proceed, and involves an analysis of the principles and procedures for the particular field of research (deMarrais & Lapan, 2004).

Research methods

Methods is a term that generally refers to the specific tools and techniques used in research (Bogdan and Biklen, 2003). Methods should be consistent with the logic embodied in the research methodology (Bogdan and Biklen). Methods are specific research tools used by researchers to gain fuller understanding of the phenomenon under investigation (deMarrais and Lapan, 2004).

This study utilized a quantitative approach, gathering publicly available, quantitative assessment data from the Texas Education Agency, to determine the extent, if any, to which state and federal accountability instruments may be impacting schools and school districts in the design and delivery of curriculum and instruction and the extent, if any, to which said accountability models disparately impact schools related to accountability-related risk factors.

The study involved a multi-grade, review of data measuring student performance on the Texas Assessment of Knowledge and Skills (TAKS) at the item-level, to identify possible correlation between a school's accountability risk and the item-level assessment performance of students within the school(s) which might suggest distortion related to the design and delivery of curriculum and instruction to maximize performance of focus or bubble students in order to best meet the increasingly pressing requirements of state and federal accountability instruments, specifically Texas's AEIS and the federal AYP provisions of NCLB.

Ethical Considerations

This researcher sought and was granted approval to conduct the study from the Institutional Review Board of the University of Texas at El Paso. The data used for the comparison of risk group item-level performance involved the use of campus-level, publicly available data, which did not

contain information about individual students or the performance of individuals from the campuses. No individual student level data was utilized or sought. All data was secondary performance level data related to the performance of campuses at the item level.

Subjects and Selection of Subjects

At the present time, more than 90 percent of the nearly 53 million children attending elementary and secondary schools in the United States are enrolled in public schools subject to federal accountability under NCLB. In Texas alone nearly 4.6 million students are enrolled in schools subject to both federal AYP and state AEIS accountability. In the selected district, more than 60,000 students subject to both AEIS and AYP accountability requirements.

The subjects of this research study were elementary public schools in two Texas public school districts; one a border area district (Far West Texas, ISD) and the other a central area district (Central Texas, ISD). The study utilized metrics collected related to item-level performance of students as well as test-score metric distributions. Fifteen, randomly selected (within risk groups) elementary schools from each focus district are included in the study. A careful examination of the schools' past accountability performance as well as socioeconomic and demographic data was done, and, based on such analyses, schools were placed into three accountability risk groups: high-risk, moderate-risk, and low-risk.

Campuses were placed into accountability exposure groups based on the following criteria:

High Risk – The high risk campus group includes all campuses that either (1) have received an accountability rating of "unacceptable" or "missed AYP" during the past five years; or (2) utilized exceptions for safe harbor, confidence intervals, subgroup sample size, or projected growth in order to meet minimum expectations for AEIS or AYP.

Moderate Risk – The moderate risk campus group consists of campuses that either (1) have received an accountability rating of "unacceptable" or "missed AYP" since the inception of NCLB AYP provisions; (2) utilized exceptions for safe harbor, confidence intervals, subgroup sample size, or projected growth, in order to meet minimum expectations for AEIS/AYP; or (3) would have received a negative accountability rating or would have needed exceptions as described above to avoid such rating using the following year's AMO targets (e.g., by applying 2010 requirements to a school's 2009 data).

Low Risk – The low risk campus group consists of campuses that (1) have not been rated "unacceptable" or "missed AYP"; (2) have not needed exceptions to achieve acceptable ratings; and (3) would not have received negative accountability ratings even with the application of the following year's AMO targets.

The review of literature pertaining to the impact of accountability requirements to schools and districts strongly suggests a difference in the impact of federal and/or state accountability for campus related to their accountability risk, particularly as such risk reflects mandated instructional practices related to the application of various levels of sanctions under the AYP accountability instrument (Rouse, et al, 2007).

Research Design

To answer the first question of this study, the researcher conducted a comparison of item-level data for schools identified at high risk of facing accountability sanctions as defined below with state item-level data to determine if the results of the focus campuses suggest a correlation between student performance and possible distortion of instructional focus. The researcher further conducted a comparison of district schools in various levels or groups of risk, ranging from high-risk, moderate risk, and low risk to determine whether such distortion, if any, can be

correlated to increasing levels of accountability risk. Campus were placed in risk groups based on past accountability performance as described above.

Both studies compared student performance of campuses in the identified accountability exposure groups on individual items ranked on a scale from easiest to most difficult (as determined by state-level item metrics (Rasch, B-equate item difficulty measure) to determine if campuses performance on items of various difficulty (i.e., requiring different levels of skill and higher-order thinking) remains consistent with the focus group or lags in areas that suggest a distortion in instructional focus.

The researcher used a simple regression model for exploring the relationship of the predictive variable, item difficulty, as defined by the state metric described above, and the dependent variable of this quantitative study, the item-level performance gap between students at campuses within each focus group. The predictive variable for the item level data will describe data from the 2010-2011 school year related to student performance on the TAKS reading and mathematics tests at grades 3 and 5. Grades three and five were chosen as these grades are part of the state's Student Success Initiative (SSI), and thus more sensitive to the demands of accountability. The SSI provisions, for example, provide for the potential retention of students who do not score proficient on the state's summative assessment. The predictive variable for accountability exposure described campuses' risk factors as described herein.

The research chose to focus the study on elementary campuses from a large west Texas border area district and a large central Texas district because the researcher believed that the diversity of these campuses would demonstrate that accountability risk factors have a distorting effect on instructional practice that is not dependent upon the presence or lack of socioeconomic factors. It was the researcher's belief that, even when controlling for the direct effects of these

factors, accountability exposure though itself indirectly related, would prove predictive to poorer student performance at the item-level, despite schools' otherwise acceptable, recognized, or even exemplary, accountability ratings.

The research design of the study was a non-experimental, explanatory, correlation design (Keppel & Zedeck, 1989) that employed simple regression analysis to measure the relationships of the predictive variable (item difficult) and the dependent variable of item-level student achievement gap on the Texas Assessment of Knowledge and Skills for the 2010-2011 school year. Non-experimental research is considered "an important and appropriate mode of research in education" (Johnson, 2001, p. 3) since such studies frequently involve areas where neither randomized experiments nor quasi-experimental designs are possible. Johnson (2001) further advises that explanatory studies must (a) develop or test theories about phenomenon that attempt to explain "how" or "why" the phenomenon occurs, and (b) tries to identify correlations that may represent potential causal factors.

Data was acquired from the Texas Education Agency technical digest and reports, TEA school report cards 2002/2003 through 2010/2011 school years for purpose of assigning campuses to accountability risk groups and from the 2010 TAKS performance data files for the purpose of determining performance levels and performance gaps. Data pertaining to the gaps in student performance were be entered into and manipulated/analyzed using SPSS/PAWS statistical software. The researcher then employed a simple regression analysis which allows correlation of the dependant variable or outcome "based on values of the predictive variables." (Field, 2009, p. 198). The use of existing data mitigated the potential reliability threat often associated with independent data gathering techniques (Suskie, 1996).

Data Analysis

The two research questions were examined by first conducting a descriptive correlational analysis to discover if the predictor variable contributes to the independent variable. The researcher utilized the following simple regression equation:

$y = \alpha + \beta x + \varepsilon$, where x represents the predictor variable (item difficulty) and y represents the outcome or dependent variable (item-level student achievement gap between the various risk groups). According to Gelman and Hill (2007), linear regression is appropriate when seeking to determine the relationship between a quantitative outcome and a quantitative explanatory variable. In this study, the research sought to determine if a significant relationship existed between the mean gap in student performance, measured item to item, and the difficulty level of the item as determined by state item response theory (IRT) parameters, specifically the Rasch differential.

To determine whether or not accountability risk was a significant factor in curricular instruction, the mean p-values for students in all schools at each of the risk levels was calculated for each item and regressed to determine a predictive relationship with regard to item difficulty, with the underlying assumption being that an increased performance gap on more difficult items would support the theory that schools subjected to escalating levels of accountability risk target instruction away from the higher-order skills necessary to successfully answer such items (even comparing the top performing students in each risk group). The null hypothesis was that no significant difference existed between the gap in student performance related to item difficulty or, in other words that students at the various ability levels within each risk group would exhibit similar achievement gaps across the spectrum of item difficulty.

Chapter Summary

This chapter began with a brief overview of the study that included a review of the purpose of the study and the research questions. The researcher also described the investigative mode and explained with the model was selected for the study. The chapter also details the ethical considerations that the researcher reviewed related to the selection of subjects, along with a description of how the researcher selected the subjects and determined their level of risk. In addition, the research design and the plan for analyzing the data were reviewed.

Looking ahead, Chapter 4 will describe the results of the statistical testing and preliminary analyses as to the significance of the independent variable as a predictor of the student achievement gap under the model. The chapter will also include a break down of each of the tests run for each of the two focus groups (limited English and economically disadvantaged) isolated and excluded from students not coded in either of those groups. Finally, the chapter will show the results of the TAAS control group testing along with a preliminary description of the data.

Chapter 4

RESULTS

This chapter will detail the preliminary results of the statistical testing for each test and student group (e.g., economically disadvantaged and non economically disadvantaged) along with a brief description related to the significance of the independent variable as a predictor of student achievement gap and the relative strength of the model. The results of the data tables will illustrate the statistical tests that were performed for each of the two focus groups (limited English and economically disadvantaged). As will be seen, separate tests for each subgroup and non-subgroup were isolated and excluded from other student groups in order to control for characteristics of the subgroup with the model. Finally, the chapter will describe the result of the TAAS control group testing along with a preliminary description of that data.

Investigative Model

Linear regression was utilized to test the hypothesis that the gap in student performance at the item level between campus groups at different risk levels could be predicted from the independent variable of item difficulty. Overall, as reported below, the model proved significant for all 2011 TAKS assessments observed, in reading and in mathematics, both at third and fifth grades, and when Limited English Proficient (LEP) and Economically Disadvantaged (ED) students were isolated in the model or excluded from the model, Non Limited English Proficient (NLEP) and Non Economically Disadvantaged (NED).

Results of the Statistical Tests

Table #1

Grade 3 Reading Low Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,40)	Sig
ED	.936	.875	.872	3.5665	281.032	.000
LEP	.928	.862	.858	3.7419	249.411	.000
NED	.802	.643	.634	4.0234	71.966	.000
NLEP	.828	.686	.678	3.6481	87.208	.000

As Table 1 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the low and high risk group on the Grade 3 reading test items relative to item difficulty, about 88 percent for the ED group and about 86 percent for the LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (63 percent) and the NLEP group (68 percent).

As is illustrated in Table 2, below, the relationship was significant for all groups In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and high risk groups likewise increases.

Table # 2

Grade 3 Reading Low Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	11.614	.693	.936	16.764	.000
LEP	11.479	.727	.928	15.793	.000
NED	6.632	.782	.802	8.485	.000
NLEP	6.618	.709	.828	9.339	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about an 11.6 unit increase in the student performance gap.

Table #3

Grade 3 Reading Low Risk-Moderate Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,40)	Sig
ED	.967	.934	.933	1.4819	569.790	.000
LEP	.989	.942	.943	1.4993	568.974	.000
NED	.718	.515	.503	3.3596	42.489	.000
NLEP	.781	.610	.600	2.8443	62.518	.000

As Table 3 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the low and moderate risk groups on the Grade 3 reading test items relative to item difficulty, about 93 percent for the ED group and about 94 percent for

the LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (51 percent) and the NLEP group (61 percent).

As is illustrated in Table 4, below, the relationship was significant for all groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and moderate risk groups likewise increases.

Table # 4

Grade 3 Reading Low Risk-Moderate Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	6.872	.288	.967	23.870	.000
LEP	6.947	.291	.930	23.753	.000
NED	4.254	.653	.718	6.515	.000
NLEP	4.369	.553	.781	7.907	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 6.9 unit increase in the student performance gap.

Table #5

Grade 3 Reading Moderate Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,40)	Sig
ED	.961	.923	.921	.9732	477.615	.000
LEP	.946	.896	.893	1.1350	343.826	.000
NED	.674	.454	.440	2.1229	33.611	.000
NLEP	.709	.502	.490	1.9793	40.318	.000

As Table 5 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the moderate and high risk group on the Grade 3 reading test items relative to item difficulty, about 92 percent for the ED group and about 90 percent for the LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (45 percent) and the NLEP group (50 percent).

As is illustrated in Table 6, below, the relationship was significant for all groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the moderate and high risk groups likewise increases.

Table # 6

Grade 3 Reading Moderate Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	4.132	.189	.961	21.854	.000
LEP	4.088	.220	.946	18.543	.000
NED	2.378	.412	.674	5.766	.000
NLEP	2.441	.384	.709	6.350	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 4.1 unit increase in the student performance gap.

Table #7

Grade 3 Math Low Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,42)	Sig
ED	.910	.828	.824	3.3633	202.501	.000
LEP	.946	.895	.892	2.6596	357.800	.000
NED	.927	.859	.855	2.9026	255.539	.000
NLEP	.916	.840	.836	3.254	220.141	.000

As Table 7 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the low and high risk group on the Grade 3 math test items relative to item difficulty, about 83 percent for the ED group and about 90 percent for the LEP

group. The model also accounted for a sizeable percentage of the variance within the NED group (86 percent) and the NLEP group (84 percent).

As is illustrated in Table 8, below, the relationship was significant for all groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and high risk groups likewise increases.

Table # 8

Grade 3 Math Low Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	8.561	.602	.910	14.320	.000
LEP	8.999	.476	.946	18.916	.000
NED	8.300	.519	.927	15.986	.000
NLEP	8.638	.582	.916	14.837	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about an 8.6 unit increase in the student performance gap.

Table #9

Grade 3 Math Low Risk-Moderate Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,42)	Sig
ED	.876	.767	.761	2.2412	138.130	.000
LEP	.910	.828	.824	1.8984	202.188	.000
NED	.877	.769	.764	2.3266	139.922	.000
NLEP	.889	.791	.786	2.1509	158.998	.000

As Table 9 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the low and moderate risk groups on the Grade 3 math test items relative to item difficulty, about 77 percent for the ED group and about 83 percent for the LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (77 percent) and the NLEP group (79 percent).

As is illustrated in Table 10, below, the relationship was significant for all groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and moderate risk groups likewise increases.

Table # 10

Grade 3 Math Low Risk-Moderate Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	4.712	.401	.876	11.753	.000
LEP	4.829	.340	.910	14.219	.000
NED	4.923	.416	.877	11.829	.000
NLEP	4.369	.553	.781	7.907	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 4.7 unit increase in the student performance gap.

Table #11

Grade 3 Math Moderate Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,42)	Sig
ED	.674	.455	.442	1.8428	35.004	.000
LEP	.702	.493	.481	1.7475	40.828	.000
NED	.767	.588	.578	1.7017	59.845	.000
NLEP	.757	.573	.562	1.6933	56.271	.000

As Table 11 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the moderate and high risk group on the Grade 3 math test items relative to item difficulty, about 46 percent for the ED group and about 48 percent for the

LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (59 percent) and the NLEP group (57 percent).

As is illustrated in Table 12, below, the relationship was significant for all groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the moderate and high risk groups likewise increases.

Table # 12

Grade 3 Math Moderate Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	1.950	.330	.674	5.916	.000
LEP	1.997	.313	.702	6.390	.000
NED	2.355	.304	.767	7.736	.000
NLEP	2.272	.303	.757	7.501	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 1.95 unit increase in the student performance gap.

Table #13

Grade 5 Reading Low Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,46)	Sig
ED	.948	.900	.897	3.5119	412.249	.000
LEP	.931	.867	.865	3.9953	301.095	.000
NED	.876	.768	.763	4.8953	152.368	.000
NLEP	.886	.786	.781	4.7017	168.474	.000

As Table 13 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the low and high risk group on the Grade 5 reading test items relative to item difficulty, about 88 percent for the ED group and about 86 percent for the LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (63 percent) and the NLEP group (68 percent).

As is illustrated in Table 14, below, the relationship was significant for all groups In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and high risk groups likewise increases.

Table # 14

Grade 5 Reading Low Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	11.077	.546	.948	20.304	.000
LEP	10.770	.621	.931	17.352	.000
NED	9.387	.760	.876	12.344	.000
NLEP	9.480	.730	.886	12.980	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about an 11.6 unit increase in the student performance gap.

Table # 15

Grade 5 Reading Low Risk-Moderate Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,46)	Sig
ED	.974	.949	.948	1.5296	852.810	.000
LEP	.962	.926	.925	1.8787	578.090	.000
NED	.907	.824	.820	2.8214	214.624	.000
NLEP	.936	.875	.872	2.3924	322.556	.000

As Table 15 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the low and moderate risk groups on the Grade 5 reading test items relative to item difficulty, about 93 percent for the ED group and about 94 percent for

the LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (51 percent) and the NLEP group (61 percent).

As is illustrated in Table 16, below, the relationship was significant for all groups In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and moderate risk groups likewise increases.

Table # 16

Grade 5 Reading Low Risk-Moderate Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	6.939	.238	.974	29.203	.000
LEP	7.017	.292	.962	24.043	.000
NED	6.421	.438	.907	14.650	.000
NLEP	6.675	.372	.936	17.960	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 6.9 unit increase in the student performance gap.

Table # 17

Grade 5 Reading Moderate Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,46)	Sig
ED	.971	.943	.941	.9633	756.320	.000
LEP	.934	.871	.869	1.5099	311.954	.000
NED	.894	.799	.795	1.7796	182.742	.000
NLEP	.922	.850	.847	1.5765	260.917	.000

As Table 17 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the moderate and high risk group on the Grade 5 reading test items relative to item difficulty, about 92 percent for the ED group and about 90 percent for the LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (45 percent) and the NLEP group (50 percent).

As is illustrated in Table 18, below, the relationship was significant for all groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the moderate and high risk groups likewise increases.

Table # 18

Grade 5 Reading Moderate Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	4.115	.150	.971	27.50A	.000
LEP	4.143	.235	.934	17.662	.000
NED	3.737	.276	.894	13.518	.000
NLEP	3.956	.245	.922	16.153	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 4.1 unit increase in the student performance gap.

Table # 19

Grade 5 Math Low Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,47)	Sig
ED	.882	.779	.774	4.5247	168.840	.000
LEP	.944	.892	.889	3.2056	394.970	.000
NED	.865	.749	.744	4.9433	143.228	.000
NLEP	.871	.759	.754	4.8202	151.244	.000

As Table 19 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the low and high risk group on the Grade 5 math test items relative to item difficulty, about 78 percent for the ED group and about 89 percent for the LEP

group. The model also accounted for a sizeable percentage of the variance within the NED group (75 percent) and the NLEP group (76 percent).

As is illustrated in Table 20, below, the relationship was significant for all groups In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and high risk groups likewise increases.

Table # 20

Grade 5 Math Low Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	8.695	.669	.882	12.994	.000
LEP	9.422	.474	.944	19.874	.000
NED	8.749	.731	.865	11.968	.000
NLEP	8.767	.713	.871	12.298	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about an 8.7 unit increase in the student performance gap.

Table #21

Grade 5 Math Low Risk-Moderate Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,48)	Sig
ED	.902	.814	.810	2.9917	210.489	.000
LEP	.942	.887	.884	2.1110	376.099	.000
NED	.887	.788	.783	3.1199	177.947	.000
NLEP	.891	.793	.789	3.1692	183.970	.000

As Table 21 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the low and moderate risk groups on the Grade 5 math test items relative to item difficulty, about 81 percent for the ED group and about 89 percent for the LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (79 percent) and the NLEP group (79 percent).

As is illustrated in Table 22, below, the relationship was significant for all groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and moderate risk groups likewise increases.

Table # 22

Grade 5 Math Low Risk-Moderate Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	6.419	.442	.902	14.508	.000
LEP	6.055	.312	.942	19.393	.000
NED	6.352	.476	.887	13.340	.000
NLEP	6.357	.469	.891	13.564	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 6.4 unit increase in the student performance gap.

Table # 23

Grade 5 Math Moderate Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,48)	Sig
ED	.920	.846	.843	1.6110	246.443	.000
LEP	.961	.923	.921	1.1850	573.165	.000
NED	.849	.721	.715	2.3419	123.942	.000
NLEP	.828	.686	.679	2.3761	104.703	.000

As Table 23 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the moderate and high risk group on the Grade 5 math test items relative to item difficulty, about 85 percent for the ED group and about 92 percent for the

LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (72 percent) and the NLEP group (69 percent).

As is illustrated in Table 24, below, the relationship was significant for all groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the moderate and high risk groups likewise increases.

Table # 24

Grade 5 Math Moderate Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	3.876	.238	.920	16.268	.000
LEP	4.196	.175	.961	23.941	.000
NED	3.856	.346	.849	11.133	.000
NLEP	3.596	.351	.828	10.232	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 3.9 unit increase in the student performance gap.

2002 TAAS Control Group Analyses

In addition to the 2011 TAKS data made the basis for the analyses described by the previous 24 tables, the study examined 2002 TAAS data. The 2001-2002 school year marked the final administration of the Texas Assessment of Academic Skills (TAAS) before it was replaced the following year by TAKS and prior to the full implementation of NCLB Accountability requirements. The researcher, therefore, hypothesized that the TAAS data would show an insignificant or less significant gap among performance levels for schools in various risk groups. A gap analysis of the TAAS student performance data for schools in the various risk

groups revealed, in most cases, a significant but weak (in comparison with those demonstrated by comparison of TAKS data) relationship between schools in the various risk groups. In other cases, the regression revealed no relationship, and in a few instances, a fairly strong relationship was determined.

Such a relationship, however, as noted earlier, was not unexpected. Although, as noted above, NCLB Accountability mandates had yet to be implemented and no schools in the study had yet been subjected to AYP requirements, let alone sanctions, the schools were subject to accountability provisions of the state system, AEIS. Like AYP under NCLB, the AEIS system relies on state assessment data to determine accountability ratings. However, the AEIS does not employ a system of escalating sanctions as does the federal system, nor did the AEIS in 2002 have in place rigorous targets for focused subgroups. The presence of an accountability system which applied pressure to educators to meet accountability guidelines is consistent with the finding of a generally weaker though significant relationship between the student performance data and item difficulty on the TAAS.

Table # 25

Grade 3 TAAS Reading Low Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,40)	Sig
ED	.424	.180	.159	0.7991	8.767	.005
LEP	.444	.197	.177	0.5662	9.794	.003
NED	.313	.098	.075	0.6743	4.332	.044
NLEP	.767	.589	.579	0.3547	57.292	.000

As Table 25 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a small percentage of the

variance in performance gap between the low and high risk group on the Grade 3 reading test items relative to item difficulty, about 18 percent for the ED group and about 20 percent for the LEP group. The model also accounted for a modest percentage of the variance within the NED group (10 percent) and a sizeable percentage of the NLEP group (59 percent).

As is illustrated in Table 26, below, the relationship was significant for all groups In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and high risk groups likewise increases.

Table # 26

Grade 3 TAAS Reading Low Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	0.455	.154	.424	2.961	.005
LEP	0.341	.109	.444	3.130	.003
NED	0.270	.130	.313	2.081	.044
NLEP	0.516	.068	.767	7.569	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a .45 unit increase in the student performance gap.

Table # 27

Grade 3 TAAS Reading Low Risk-Moderate Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,40)	Sig
ED	.541	.292	.274	0.5753	16.511	.000
LEP	.572	.327	.310	0.5363	19.419	.000
NED	.573	.329	.312	0.5355	19.599	.000
NLEP	.590	.348	.331	0.5301	21.305	.000

As Table 27 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a modest percentage of the variance in performance gap between the low and moderate risk groups on the Grade 3 reading test items relative to item difficulty, about 29 percent for the ED group and about 33 percent for the LEP group. The model also accounted for a modest percentage of the variance within the NED group (33 percent) and the NLEP group (35 percent).

As is illustrated in Table 28, below, the relationship was significant for all groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and moderate risk groups likewise increases.

Table # 28

Grade 3 TAAS Reading Low Risk-Moderate Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	0.449	.111	.541	4.063	.000
LEP	0.454	.103	.572	4.407	.000
NED	0.456	.103	.573	4.427	.000
NLEP	0.470	.102	.590	4.616	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 0.45 unit increase in the student performance gap.

Table # 29

Grade 3 TAAS Reading Moderate Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,40)	Sig
ED	.424	.180	.159	0.7991	8.767	.005
LEP	.364	.132	.111	0.8077	6.095	.018
NED	.493	.243	.224	0.7382	12.831	.001
NLEP	.245	.060	.036	0.8330	2.545	.119

As Table 29 illustrates, the F-test indicates the regression model is statistically significant for all groups with the exception of the non-LEP group, which is not significant. However, the R Square value reflects that the model accounts for a modest percentage of the variance in performance gap between the moderate and high risk group on the Grade 3 reading test items relative to item difficulty, about 18 percent for the ED group and about 13 percent for the LEP

group. The model also accounted for a modest percentage of the variance within the NED group (24 percent).

As is illustrated in Table 30, below, the relationship was significant for all groups except the non-LEP group. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the moderate and high risk groups likewise increases.

Table # 30

Grade 3 TAAS Reading Moderate Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	0.455	.154	.424	2.961	.005
LEP	0.383	.155	.364	2.469	.018
NED	0.508	.142	.493	3.582	.001
NLEP	0.255	.160	.245	1.595	.119

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 0.46 unit increase in the student performance gap.

Table # 31

Grade 3 TAAS Math Low Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,44)	Sig
ED	.664	.442	.428	0.5631	33.209	.000
LEP	.525	.275	.258	0.6944	15.940	.000
NED	.634	.401	.387	0.5352	28.175	.000
NLEP	.630	.397	.383	0.6095	27.667	.000

As Table 31 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a moderate percentage of the variance in performance gap between the low and high risk group on the Grade 3 math test items relative to item difficulty, about 44 percent for the ED group and about 28 percent for the LEP group. The model also accounted for a moderate percentage of the variance within the NED group (40 percent) and the NLEP group (40 percent).

As is illustrated in Table 32, below, the relationship was significant for all groups In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and high risk groups likewise increases.

Table # 32

Grade 3 TAAS Math Low Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	0.564	.098	.664	5.763	.000
LEP	0.451	.113	.525	3.992	.000
NED	0.494	.093	.634	5.308	.000
NLEP	0.558	.106	.630	5.260	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about an 0.56 unit increase in the student performance gap.

Table # 33

Grade 3 TAAS Math Low Risk-Moderate Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,44)	Sig
ED	.416	.173	.153	0.8090	8.770	.000
LEP	.572	.327	.310	0.5363	19.419	.000
NED	.684	.468	.456	0.5230	36.996	.000
NLEP	.288	.083	.061	0.8118	3.788	.058

As Table 33 illustrates, the F-test indicates the regression model is statistically significant for all groups, except the non-LEP group, which is not significant, though barely. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the low and moderate risk groups on the Grade 3 math test items relative to item difficulty, about 17 percent for the ED group and about 33 percent for the LEP group. The

model also accounted for a moderate percentage of the variance within the NED group (46 percent).

As is illustrated in Table 34, below, the relationship was significant for all groups, except the non-LEP group. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and moderate risk groups likewise increases.

Table # 34

Grade 3 TAAS Math Low Risk-Moderate Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	0.417	.141	.416	2.961	.005
LEP	0.454	.103	.572	4.407	.000
NED	0.553	.091	.684	6.082	.000
NLEP	0.275	.141	.288	1.946	.058

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 0.41 unit increase in the student performance gap.

Table #35

Grade 3 TAAS Math Moderate Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,44)	Sig
ED	.424	.180	.159	0.7991	8.767	.005
LEP	.364	.132	.111	0.8077	6.095	.018
NED	.899	.808	.804	0.3270	177.170	.000
NLEP	.824	.679	.671	0.3850	88.851	.000

As Table 11 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a moderate percentage of the variance in performance gap between the moderate and high risk group on the Grade 3 math test items relative to item difficulty, about 42 percent for the ED group and about 36 percent for the LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (81 percent) and the NLEP group (68 percent).

As is illustrated in Table 36, below, the relationship was significant for all groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the moderate and high risk groups likewise increases.

Table # 36

Grade 3 TAAS Math Moderate Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	0.455	.154	.424	2.961	.005
LEP	0.383	.155	.364	2.469	.018
NED	0.757	.057	.899	13.311	.000
NLEP	0.631	.067	.824	9.426	.000

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 0.45 unit increase in the student performance gap.

Table # 37

Grade 5 TAAS Reading Low Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,44)	Sig
ED	.353	.125	.106	0.6582	6.545	.014
LEP	.506	.256	.240	0.6267	15.843	.000
NED	.444	.198	.180	0.6454	11.322	.002
NLEP	.310	.096	.076	0.6459	4.876	.032

As Table 37 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a modest percentage of the variance in performance gap between the low and high risk group on the Grade 5 reading test items relative to item difficulty, about 13 percent for the ED group and about 26 percent for the LEP group. The model also accounted for a moderate percentage of the variance within the NED group (20 percent) and the NLEP group (10 percent).

As is illustrated in Table 38, below, the relationship was significant for all groups In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and high risk groups likewise increases.

Table # 38

Grade 5 TAAS Reading Low Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	0.267	.104	.353	2.558	.014
LEP	0.396	.099	.506	3.980	.000
NED	0.344	.102	.444	3.365	.002
NLEP	0.226	.102	.310	2.208	.032

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 0.27 unit increase in the student performance gap.

Table # 39

Grade 5 TAAS Reading Low Risk-Moderate Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,46)	Sig
ED	.027	.001	-.021	0.6692	0.0330	.857
LEP	.226	.051	.030	0.6928	2.4750	.123
NED	.113	.013	-.009	0.6987	0.5920	.446
NLEP	.003	.000	-.022	0.6489	0.0000	.986

As Table 39 illustrates, the F-test indicates the regression model is not statistically significant for any group. The R Square value reflects that the model accounts for less than a significant percentage of the variance in performance gap between the low and moderate risk groups on the Grade 5 reading test items relative to item difficulty.

As is illustrated in Table 40, below, the relationship was significant for only the LEP subgroup. In addition, though the coefficients for all groups except the non-LEP subgroup are

positive, indicating that as item difficulty increases the performance gap between campuses in the low and moderate risk groups likewise increases, the strength of the association is extremely weak, and therefore unreliable, which conclusion is reinforced by the negative coefficient for the non-LEP subgroup score, indicating, counter intuitively, that as item difficulty increases the performance gap decreases.

Table # 40

Grade 5 TAAS Reading Low Risk-Moderate Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	0.019	.106	.027	0.181	.857
LEP	0.173	.110	.226	1.573	.000
NED	0.085	.111	.113	0.769	.446
NLEP	- 0.002	.103	- .003	- .0180	.986

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the LEP group that each one unit increase in difficulty yields about a .17 unit increase in the student performance gap.

Table # 41

Grade 5 TAAS Reading Moderate Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,46)	Sig
ED	.457	.209	.191	0.4918	12.120	.001
LEP	.634	.402	.389	0.4633	30.899	.000
NED	.569	.323	.309	0.4674	21.983	.000
NLEP	.477	.228	.211	0.4996	13.573	.001

As Table 41 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a modest percentage of the variance in performance gap between the moderate and high risk group on the Grade 5 reading test items relative to item difficulty, about 21 percent for the ED group and about 40 percent for the LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (32 percent) and the NLEP group (23 percent).

As is illustrated in Table 42, below, the relationship was significant for all groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the moderate and high risk groups likewise increases.

Table # 42

Grade 5 TAAS Reading Moderate Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	0.271	.078	.457	3.481	.001
LEP	0.408	.073	.634	5.559	.000
NED	0.348	.074	.569	4.689	.000
NLEP	0.292	.079	.477	3.684	.001

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about a 0.27 unit increase in the student performance gap.

Table # 43

Grade 5 TAAS Math Low Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,48)	Sig
ED	.353	.125	.106	0.6582	6.545	.014
LEP	.394	.155	.138	0.6843	8.826	.005
NED	.393	.154	.137	0.6902	8.766	.005
NLEP	.310	.096	.076	0.6458	4.876	.032

As Table 43 illustrates, the F-test indicates the regression model is statistically significant for all groups. The R Square value reflects that the model accounts for a sizeable percentage of the variance in performance gap between the low and high risk group on the Grade 5 math test items relative to item difficulty, about 13 percent for the ED group and about 16 percent for the LEP group. The model also accounted for a sizeable percentage of the variance within the NED group (15 percent) and the NLEP group (10 percent).

As is illustrated in Table 44, below, the relationship was significant for all groups In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and high risk groups likewise increases.

Table # 44

Grade 5 TAAS Math Low Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	0.267	.104	.353	2.558	.014
LEP	0.301	.101	.394	2.971	.005
NED	0.302	.102	.393	2.961	.005
NLEP	0.226	.102	.310	2.208	.032

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the ED group that each one unit increase in difficulty yields about an 0.27 unit increase in the student performance gap.

Table # 45

Grade 5 TAAS Math Low Risk-Moderate Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,48)	Sig
ED	.027	.001	- .021	0.6692	0.033	.857
LEP	.182	.033	.013	0.8650	1.637	.207
NED	.364	.133	.115	0.6523	7.338	.009
NLEP	.387	.149	.122	0.6989	7.870	.015

As Table 45 illustrates, the F-test indicates the regression model is statistically significant for only the NED and NLEP groups. The R Square value reflects that the model accounts for a small percentage of the variance in performance gap between the low and moderate risk groups on the Grade 5 math test items relative to item difficulty, about 36 percent for the NED group and 39 percent for the NLEP group.

As is illustrated in Table 46, below, the relationship was not significant for the ED and LEP groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the low and moderate risk groups likewise increases.

Table # 46

Grade 5 TAAS Math Low Risk-Moderate Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	0.019	.106	.027	0.181	.857
LEP	0.164	.128	.182	1.280	.207
NED	0.262	.097	.364	2.709	.009
NLEP	0.369	.153	.481	2.907	.015

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the NLEP group that each one unit increase in difficulty yields about a 0.37 unit increase in the student performance gap.

Table # 47

Grade 5 TAAS Math Moderate Risk-High Risk Gap Summary

Group	R	R Square	Adjusted R Square	Standard Error of the Estimate	F(1,42)	Sig
ED	.457	.209	.191	0.4912	12.120	.001
LEP	.283	.080	.061	1.0084	4.181	.046
NED	.609	.371	.358	0.6041	28.275	.000
NLEP	.477	.228	.211	0.4996	13.573	.001

As Table 47 illustrates, the F-test indicates the regression model is statistically significant for all groups, though only slightly so for the LEP group. The R Square value reflects that the model accounts for a modest percentage of the variance in performance gap between the moderate and high risk group on the Grade 5 math test items relative to item difficulty, about 21 percent for the

ED group and about 8 percent for the LEP group. The model also accounted for a moderate percentage of the variance within the NED group (37 percent) and the NLEP group (23 percent).

As is illustrated in Table 48, below, the relationship was significant for all groups. In addition, the coefficients for all groups are positive, indicating that as item difficulty increases the performance gap between campuses in the moderate and high risk groups likewise increases.

Table # 48

Grade 5 TAAS Math Moderate Risk-High Risk Gap Analysis

Group	B	Std. Error	Beta	t	Sig
ED	0.271	.078	.457	3.481	.001
LEP	0.305	.149	.283	2.045	.046
NED	0.475	.089	.609	5.317	.000
NLEP	0.292	.079	.477	3.684	.001

As the table indicates, the sizes of the coefficients vary, but indicate, for example in the NED group that each one unit increase in difficulty yields about a 0.48 unit increase in the student performance gap.

Chapter Summary

This chapter described the preliminary results of the statistical testing for each test and student group (e.g., limited English proficient, non-limited English proficient, economically disadvantaged, and non economically disadvantaged), provided a brief discussion as to the significance of the independent variable as a predictor of student achievement gap, and described the relative strength of the model. The results of analyses were reported in tables that illustrated the statistical tests that were performed for each of the focus groups (limited English proficient and economically disadvantaged/non-limited English proficient and non economically

disadvantaged). The results reported separate tests for each subgroup and non-subgroup, isolated and excluded from other student groups, to control for subgroup characteristics not accounted for in the basic model. Lastly, the chapter describes the results of the TAAS control group testing along with a preliminary description of that data.

The final chapter will describe the study context, a review of the study questions, and the methods the researcher used to conduct the study, as well as a description of the results and how they relate to the theoretical framework. The researcher includes some preliminary conclusions drawn from the data along with some suggestions for further research related to the study. The chapter then reviews the extant literature in light of the preliminary study results. Finally the chapter describes possible implications for both practice and policymakers, and ends with some personal, concluding remarks.

Chapter 5

DISCUSSION

This chapter will describe the context of the study, briefly review study questions as well the methods employed by the researcher for the study, and discussion of theoretical framework as it relates to the results. Next the researcher draws a few brief preliminary conclusions based on analyses of the results and describes some suggestions for further research. A review of the extant literature follows, adding the perspective of the study results. The chapter ends with a description of possible implications for practice as well as policymakers, along with some final, concluding remarks.

Study Context

A major issue this study attempts to address relates to the efficacy of policies that hold school districts, schools, and teachers accountable for the performance of students on state administered summative assessments, such as the TAKS test. The evidence, as discussed in the introduction to this dissertation, is by and large inconclusive, with a variety of authors, such as Braun (2004), Carnoy and Loeb (2002), and Hanuskeck and Raymond (2004), presenting evidence that seems to suggest a link between high-stakes accountability and improved student achievement. With regard to this study, the Carnoy and Loeb (2002) study is especially provocative, in that the authors found that the greater the external pressure created by the accountability system, the greater the improvement gains in student achievement. In contrast, the this dissertation sought to establish that the external pressure placed upon schools, at least to the extent that some schools are designated as low performing or failing to meet adequate yearly progress, has a detrimental impact on students in that such pressure induces educators to design and deliver instruction in order to realize the quickest, easiest, and surest score gains on the state

assessment instrument. Although these two theories appear to be diametrically opposed, they are less so than might be assumed at first glance. Moreover, this study is not directly concerned with the raw test scores of any one group of students but rather the gap between student groups and how such gap, and changes in it, relate to the instructional philosophy practiced at the schools and whether that philosophy reflects distortion created by the pressures exerted through the accountability system.

In analyzing the results of this study, it is vital to keep in mind several implications related to the general theory of distortion and the variety of ways in which it manifests itself in public schools. First, as noted earlier in this study and should be intuitively evident, if school practices, such as the targeted instruction suggested herein, are able to mask significant problems in the instructional environment from detection, perception of which is obscured by myopic accountability system, then public policy in this realm has failed. Far worse, however, if such practices do more than passively mask existing problems, but rather systemically alter the instructional program so that it no longer serves every student to the full range of their potential, as the results of this study suggest, then the accountability instrument established to implement said policy has not only failed, it has become a far bigger problem than the one it was designed to solve. Second, when the design of the system, as detailed herein, creates large inequities between measures, expectations, standards of achievement, and proficiency targets, such a system cannot practically be utilized as an arbiter of school efficacy.

Research Questions

This research study was guided by two primary questions related to the impact of high-stakes accountability policies on the educational practices of teachers, schools, and school districts:

1. Does state assessment data support the theory that high stakes accountability systems encourage educators to disproportionately direct (distort) instructional practices to minimum skills levels?
2. Does the distortion of instructional practices, if any, increase subject to the accountability exposure of schools?

This study was conducted with data collected from two large Texas school districts, which will herein be referred to using the following pseudonyms: Far West Texas ISD and Central Texas ISD. The researcher selected five schools from each district for each category of risk, for a total of fifteen schools from each district. The total aggregate number of schools represented in the study was thirty. The thirty schools had a combined student population of approximately 2718 at grade 3 and 3364 at grade 5 who participated in the 2011 TAKS administration for reading and math, though the numbers varied slightly from reading to math in both cases. This is most likely due to students moving into and/or out of the state between administrations.

Methods

Individual student scores were not used in this study. Texas technical digest reports and data files (all publicly available) were used to determine the average score for the students at each campus on each item within the test administrations. The average for each campus was multiplied by the number of students who took part in the administration of the test at that campus. Then, the scores of all campuses were summed and divided by the total number of students in the risk group. This important step was done to alleviate the potential for schools that

had exceptionally higher or lower populations to skew the sample. From these figures, an overall average for campuses within each risk group was calculated for each item on the four tests that were studied (Grade 3 reading and math, and Grade 5 reading and math). Then, the researcher calculated the gap between the scores on each item between the high and low risk groups, the high and medium risk groups, and the medium and low risk group. The data tables were then analyzed using simple linear regression with SPSS statistical software version 20, to determine if the gap was significantly related to item difficulty (as defined by state IRT parameters, specifically the item's Rasch differential statistic).

If the gap increase is significantly related to item difficulty, then one might infer as a possible cause that students in the medium and especially high risk groups were receiving a disproportionate and inappropriate amount of instruction at the basic skills level. In all cases, the results of the analyses were that the gap was significantly and positively impacted by the difficulty of the items, meaning that the gap increased predictably with the level of increased difficulty.

Theoretical Framework

The theoretical framework for this study is based principally on Campbell (1979) and Simon's (1978) work with performance measurement and its relationship to incentive theory as developed by Laffont and Martimort (2001). Both of these concepts have relevance to the hypotheses of this study and are reflected in the performance of the students at the campus which are the subject thereof. As discussed earlier, Campbell's (1979) Law of Performance Measurement directly relates the underlying theory upon which this study was designed. Campbell's law attempts to explain the disparity between accountability goals and actual accountability outcomes and speculates that the failure of accountability systems to incentivize

actors as intended often stems from the systems' reliance on flawed metrics. Metrics can be flawed for several reasons, including indirect or loose alignment, poor reliability, sampling error, and deception.

In this dissertation several of these possible flaws have been discussed with regard to assessment metrics. The most egregious flaw concerns the alignment of the metric and the goal. Seeking to explain why public accountability systems frequently failed, whether they concerned public education or public transportation, Campbell (1979) determined that the accountability systems commonly used metrics that were poor fits for the outcomes they were intended to measure. At first glance, a test score would seem a natural match for a system designed to measured student achievement. However, much debate concerns how well standardized tests actually measure student achievement. The main reason cited by researchers, such as McNeil (2000) and Wright (2002), is that summative assessments cover only a small portion of the state curriculum. The reasons for this are principally financial and practical in nature. Many standards in the state curriculum are observational or chart something over time, elements that cannot be emulated on a standardized test, particularly in the context of multiple choice questions. Nor would such tests be financially feasible. Finally, given the already trying nature of state testing, it is scary to imagine students being subjected to the much lengthier tests that would by necessity have to be given to accommodate a test covering the full curriculum.

In response, schools, according to Madaus (1988), naturally have tended to focus on the areas being assessed. He notes that this is especially true in schools that are struggling to meet accountability demands. McNeil & Valenzuela (2000) have specifically addressed the harmful impact of testing in Texas, noting the targeting of tested curriculum not only in scope but in depth. On the other hand, proponents such as Crocker (2005) and Wilson (2002) have expressed

not only comfort, but indeed enthusiasm for the idea that testing play a central role in shaping curriculum and instructional practice. Therefore, it is hardly surprising that school districts and schools have come close to officially sanctioning such practice. For example, Wilson (2002) argues that the assessments have been useful in encouraging teachers to cover materials that they might in the past have found excuses to overlook or ignore because the material was not something they enjoyed teaching or fully understood themselves. Likewise, Crocker (2005) feels the role of assessment in determining what is taught in the classroom is appropriate, particularly when the assessments are "tied to state-developed curricular standards" (166). Although such an endorsement may sound tacitly logical, the impact on curriculum that most educators worry about is not what is getting taught so much as what is not getting taught and at what levels the instruction that does take place is delivered. As Padilla (2005) has noted, accountability pressures have shifted schools focus away from the primary goal of education into a battle to stay open. When schools begin to practice this type of educational triage and make decisions based more on accountability concerns than on the interests of a student's education, clearly the latter will suffer. That schools frequently make decisions about not only what is taught, but also to what extent it must be taught to achieve satisfactory results, helps to explain why students at those schools master higher-order thinking skills at an even more anemic rate than they do other skills as the results of this study suggest, even when the overall "passing" rate qualifies the school for academic accolades such as an "exemplary" rating or a Blue Ribbon School designation. When schools which are rated "acceptable" exhibit much stronger performance levels on the most difficult test items than school rated "commended" or "exemplary," as this analysis showed, then the metric is clearly not aligned to the goal.

In addition to alignment issues with metrics trying to measure the efficacy of public accountability systems, data reliability is also a factor that can create a flawed metric. As described herein, Linn (2009) and Koretz (2000), among others, have raised troubling questions about the validity of the test results as accurate measures, even of the tests themselves, let alone of the efficacy of a school. The fact that these researchers have found compelling evidence that much of score variance is unrelated to instructional differences but is more a function of non-persistent factors, lends credence to the idea that the test scores are poorly aligned indicators, as well as being flawed from a validation standpoint. In addition, as Rothstein (2008) has argued, targeting instruction disproportionately to certain areas likely to yield the best accountability results is a type of sampling error as is the practice of manipulating testing cohorts with the same intention. Such manipulation may not rise to the level of outright fraud which educators have increasingly been involved in, but they fit the model that Campbell (1979) and Simon (1978) envisioned related to flawed metrics and their role in corrupting the institution they are trying to measure.

The assertion that the corrupting influence of incentives takes on an added intensity when risk is introduced has been the subject of much of the work of Laffont and Martimort (2001). The researchers examined the role of risk in an incentive-based relationship and determined that the greater the risk imbalance the less effective the incentive became in affecting a desired behavior. In an incentive-based relationship one party attempts to encourage a desired behavior, in this case, a robust effort to improve student learning, by offering an incentive to reward evidence of the desired behavior. Many incentive-based relationships are deemed risk-neutral, meaning that the risk which the actor is subjected to is not increased or decreased by the

presence of the incentive. An actor failing to evidence the desired indicator would not earn the incentive reward but would not be subject to further penalty.

However, incentive-based relationships can also be risk averse, meaning that the incentive, which would more accurately be called a disincentive, subjects the actor to increased levels of risk because evidencing the desired behavior does not create a reward for the actor but rather allows a continuation of the status quo. Accordingly, Laffont and Martimort (2001) argue, risk averse systems in accordance with Campbell's (1979) theory of performance measurement tend to have a very corrupting influence on the system they are trying to measure. This tendency occurs because these systems induce actors to take shortcuts which may or may not be consistent with the goal, in an attempt to preserve the status quo. This is especially true, according to Laffont and Martimort when the indicators are considered unfair or unreasonable.

In the case of educators, trying to avoid sanctions, public humiliation, and ultimately loss of employment, such shortcuts are manifested by way of attempts to circumvent the intended processes so as to enhance the probability of achieving the desired indicator. Relating back to Figures 1 and 2 herein, then, educators may look to circumvent the desired continuous improvement model by avoiding rather than confronting challenges that threaten to impair or prevent the educators from achieving the desired indicator and preserving the status quo. The circumvention often takes the form of restricting the curriculum, teaching to the test, and even cohort manipulation. In still rare, but increasingly common, situations, the circumvention has taken the form of fraudulent activity. The contention that educators may be especially susceptible to risk averse behavior is bolstered by the perception of many that the system is unfairly applied, in that it relies on metrics that are not truly comparable from actor to actor, and that it is unreasonable because it sets an unrealistic goal of 100 percent proficiency.

This study is based on this risk-aversion premise, as it argues that schools at risk for accountability sanctions skew instruction in an effort to meet accountability requirements. Moreover, the risk that schools are subject to is not static, but rather becomes more adverse over time as campuses continue to fall short of the desired indicator. With each successive failure, additional and more punitive sanctions are imposed putting the campus at greater risk and increasing the likelihood of risk averse behavior. The results of this study, which indicate a drop off in student performance between risk groups as difficulty increases, supports such a theory.

Conclusions

Many studies that seek to understand relationships between two sets of separate observations suffer somewhat from the chicken and egg dilemma. This is nowhere more true than in the field of education, where educators are continuously seeking to understand why some students are more successful than others. In doing so, however, educators must continually be on guard to clarify whether a certain activity or behavior (e.g., participation in band or orchestra) is in fact a contributor to student achievement as a cursory glance at the data may suggest, or whether the data is merely a reflection of the types of students who choose to participate in band or orchestra. As Pearl (2009) notes, this is a Simpson's paradox, as first described more than a century ago by Karl Pearson, who held that any statistical relationship between two variables can be reversed by including additional factors in the analysis. For example, Pearl notes that a test of two groups may seem to indicate that participants who smoke score better on achievement tests than those who do not. Controlling for age, however, the researcher may discover that smoking instead predicts lower achievement. On the other hand, a further factor added to the model, for example, parental education or household income, may mitigate such results.

Consistent with Simpson's paradox then, conclusions related to student achievement based on assessment data are certainly not irreproachable. However, the observations made as part of this study combined with the substantial body of existing research related to the impact of accountability on curriculum and instruction provide reasonable support for the theory which serves as the basis for this study. The statistical models applied herein suggest a very strong relationship between the gap in students' abilities with regard to successfully answering questions requiring higher order skills and the extent to which accountability risks may be impacting instructional practice at their campuses. As Oakes (2005) has noted, students who do not have access to demanding curriculum and challenging instruction are "less likely to develop the higher-order thinking skills necessary to solve rigorous problems" (128). Consistent with Oakes' argument, a study of Texas students who performed badly on the state assessments noted a persistent lack of higher-order thinking skills necessary to correctly answer questions on the test, especially those that require reasoning and extension of ideas between concepts. (Texas Study of Students at Risk: Case Studies Supporting Ninth Graders' Success, 2004).

Clearly, student performance at campuses in the high and moderate risk group indicated a potential deficit in the higher order thinking skills required to answer the more difficult questions on the test. Certainly, it would be ill advised to suggest that such a deficit was entirely the result of how campuses in those risk groups reacted to the pressures of avoiding accountability sanctions under AYP. Likewise, however, it would be counterintuitive to insist that such practices as assessment worksheets, classes focused on assessment remediation, mock testing, and the insistence that students learn trendy test-taking strategies, would not necessarily result in a reduction of instruction targeted at the higher order skills necessary to answer even the toughest questions on the assessment, to say nothing of the difficulties students may face upon

leaving the controlled assessment environment and entering the world where multiple choice questions are few and far between.

The findings of this study suggest that students who attend campuses with significant accountability risk, may not be getting the exposure to higher order thinking which Oakes (2009) believes they need and that is consistent with the Texas study described earlier. In each of the twenty-four regression analyses performed to compare performance between campus risk groups while isolating to mitigate other potential contributing factors, such as socioeconomic status and limited English proficiency, a significant difference was found in the mean item score of the groups that widened substantially as item difficulty increased. Therefore, it is reasonable to assume, that students at the campuses in the high and moderate risk categories were less prepared to handle higher order cognitive demands than students in the low risk category. Whether or not the accountability system is creating, aggravating, or merely illuminating this problem may be a subject for continued study and debate, but one thing seems obvious: the accountability system is not alleviating such gaps as it was intended to do and as was premised as the underlying foundation upon which this unprecedented policy effort was constructed.

Relative to the research questions underlying this study, the state assessment data at the item level for both the 2011 TAKS reading and math assessments indicates a significant relationship to the item level difficulty both at grade 3 and grade 5. The relationship persisted both when LEP and ED students were isolated and factored out as part of the study. In this respect, the data is consistent with the concept that schools with substantial accountability risk may distort typical instructional practice and focus disproportionately on test preparation, though they do so at the expense of opportunities to maximize the educational development of students to their fullest potential.

To this extent, the results of the study support the theory that schools in the high and moderate risk groups are more successful at elevating students to minimum skills levels than at extending them beyond that level. Further, as the study revealed a significant effect not only between schools in the high and low risk groups, but also revealed significant effects both between the low and the moderate risk group and between the moderate and high risk group, it is reasonable to assume that the difficulty of preparing students for the more challenging and rigorous of the test questions, increases in accordance with the campuses' accountability risk.

Regarding the question of whether such distortion increases as risk levels increase, it would seem logical to infer that the difficulty of preparing students for the most challenging and rigorous test questions, as described in the paragraph above, is likely a function of the educational philosophy of the school. It is also likely that schools within the moderate and high risk groups may indeed adopt strategies and adapt practice in order to more directly address immediate concerns of sanctions and accountability, not to mention job security, at the expense of instruction that might otherwise have provided students with more opportunity to engage in and work through difficult test questions which require higher order skills.

Links to Extant Literature

The relatively strong relationship between the gaps demonstrated by students attending schools subjected to different levels of accountability risk with the difficulty of the test item and the conceptual ability needed to attack and solve the most difficult items align well with Campbell (1979) and Simon's (1978) theoretical framework regarding the tendency of accountability systems to corrupt and distort the very systems which they are trying to measure. Other studies and authors who have taken interest in the impact of accountability on schools and

education include Donovan, Figlio, and Rush (2006) whose study involved the effects that school accountability has had on college-bound students.

This study looked at the performance of students from schools that have faced accountability sanctions once they entered college to determine if study skills and educational approaches to which the students were exposed, presumably due to such accountability pressures versus the approaches students presumably would have been exposed to at schools without such concerns, to determine if there were differences in how the students faired in college and how they approached studies. One finding from the study suggested that students from schools that had faced accountability pressures were more likely to neglect studies until shortly before an assignment or exam was due and to then cram to prepare, with the implication being that this approach is a common tactic employed by schools desperate to achieve minimum proficiency marks.

I found this an especially compelling study in that it suggested a negative impact on a population of students (who later entered college) who likely were not the focus of the instructional practices they were subjected to, but who rather were impacted, at least in the authors' opinion, because they were subjected to curriculum and instructional practices for which they were ill-suited, rather than challenged and extended in such a manner that would have better prepared them for the rigorous educational demands they would face in college.

Hanushek and Raymond (2004) looked at how school accountability systems impact the level and distribution of student achievement. Their study used NAEP results and involved measuring not only gains and losses at the state level for participants from states with strict accountability systems (i.e., those that imposed some form of sanction for poor performance) against those from participants from states that had softer accountability systems that, for

example, may have only reported results from annual testing. The authors noted that the overall impact of adopting a strong accountability instrument seemed to be positive, but note that the largest effects were seen among students in states where educational achievement was poorest prior to adopting the accountability instrument and were far less pronounced in states that already enjoyed fairly strong test scores even prior to implementing a strong accountability instrument. Such findings suggest that high-stakes accountability systems are effective for raising achievement levels for low performing students but have little impact on average or higher performing students. Therefore, states that had a smaller percentage of low performing students showed less improvement, principally because the targeted instruction adopted by many schools in response to the accountability mandates, impacted fewer students. Although such instruction may have resulted in substantial gains for lower performing students, the instruction was less effective at increasing student achievement for average and higher performing students, presumably because it did not help these students improve with regard to the higher-order thinking skills necessary to answer the more difficult questions on the assessment. Hanushek and Raymond's (2004) findings are consistent with the findings of this study in that both suggest that accountability systems encourage schools to target instruction to low performing students to the detriment of students at other skills levels.

Finally, as a counter balance, Jacob (2004) made a compelling case that accountability measures had led to strong improvements in the Chicago public schools. Jacob's arguments are a reminder of the lure of accountability and why NCLB was passed in a strongly bipartisan fashion back in 2001. They also shed light on why, despite numerous problems including those cited within this study as well as others, accountability retains fairly strong public support. The data that Jacob presents seems to suggest that a flawed accountability system may be preferable

to no accountability at all. It should be noted, likewise, that every school in this study met AYP requirements in 2011. In addition, several of the schools within the high and moderate risk groups were rated as "Recognized" or "Exemplary" on the Texas AEIS system. However, as by definition all of these schools have at one time or another been subject to sanctions, many as far a level three, and yet, those schools have now improved scores sufficiently to meet AYP requirements despite the fact that these requirements have become more rigorous almost every year. This in all probability could not have happened without some significant improvement being made by many of the students at these schools.

However, despite attaining comparable, and in many cases, superior accountability ratings to schools in the low risk group, schools in the moderate and high risk group lagged behind schools in the low risk group when it came to the more difficult and challenging items on the tests. Furthermore, schools in the high risk group also lagged behind the moderate risk group, though to a less extent. In addition, the data clearly shows that the performance gap widens the more and more item difficulty increases, strongly suggesting that the instruction at campuses in the study focuses less and less on higher order thinking skills as accountability risks increased. The paradox suggested by this data in terms of percentages of students meeting proficiency requirements and the poor performance of students on the most rigorous of the test items is what this study sought to resolve. Is outcome-based accountability, to coin a phrase, a blessing or a curse?

The fact that some students have certainly benefited from the implementation of a strong accountability system, is consistent with one of the chief arguments in support of NCLB both before and since its enactment. Especially with regard to the subgroup requirements of the legislation which for the first time required educators to demonstrate achievement among all

groups of students. I strongly suspect, however, that not everyone in Chicago has benefited from strict accountability measures, nor I suspect has every student in the campuses in this study, despite their escape from accountability sanctions and their admirable state ratings. Nevertheless, I think it would be foolish to argue that accountability has not positively impacted many students, especially the very low performing students who may have been "written off," so to speak, prior to NCLB, but who have now become a primary focus.

Recommendations for Further Research

1. One of the great challenges of this study was finding a way to isolate the effects of instructional distortion within the data. Ultimately, the researcher chose to run each of analyses multiple times isolating and separating student groups so as to avoid the thorny issue of whether attributes of the student group population (e.g., limited English speakers) were responsible for the student performance trends in question rather than the accountability pressures associated with inclusion in an accountability risk group. However, it remains to be determined to what extent, such distortion, if it is in fact occurring as I believe it does, may impact such groups differently. For example, a curriculum and instructional approach that focused on minimal skills and gimmick test taking strategies to answer test questions might not provide students with limited English abilities the challenging tasks necessary to truly master the language at a level necessary to succeed in certain professional careers. It would be interesting, in this regard, to compare such students in this vein against schools from the various risk groups to see if a significant effect can be isolated.

2. It should be noted that the schools placed within the various risk categories for this study were not uniformly high or low as to school rankings as one might expect. Rather, as noted earlier in this chapter, no schools in the study missed AYP the previous year. In fact,

several schools among the high risk group achieved recognized status, with two being rated exemplary. One has since been nominated as a Blue Ribbon School. Similarly, although schools in the low risk group by definition would not have ever received a negative accountability rating under NCLB or within the past ten years on the state instrument, not all of them were recognized or exemplary. In fact, several of these schools had remained at the acceptable level for years. In addition to being an interesting commentary on the differences between the state and federal accountability system, I believe it raises several questions about how schools approach instruction. It would be interesting to look at school assessment data through the various ratings lenses to see how such rankings may have been related to future performance. For example, a researcher might track the evolution of schools that were rated acceptable versus schools that may have been relatively similar but missed meeting proficiency requirements by a small margin or at one grade level or in one subject area. Since districts often implement drastic changes when schools fail to meet minimum proficiency requirements, it would be interesting to analyze the disparate paths the previously similar schools may have taken after one school received a negative accountability rating. A longitudinal analysis of this nature could relate to theories espoused in this study as to the extent which new strategies or instructional approaches were effective in remediating deficiencies at the sanctioned school or whether such practices created a more risk averse relationship as the framework of this study and the results of the analysis thereof suggest is might.

3. As noted earlier in this chapter, risk associated with failing to meet accountability requirements under AYP is not static but continues to escalate until schools demonstrate sufficient improvement to meet these requirements. In fact, schools previously subject to sanctions must meet all AYP guidelines for two consecutive years before being released entirely

from sanctions. So, though this study looked at schools in three broad categories of risk, the risk continuum for campuses under AYP is far more elaborate and can lead eventually to re-staffing and even closing of a school that continues to fall short of accountability requirements for several consecutive years. A study that analyzed schools as they progressed through the various levels of sanctions to determine if the schools, as the theories of Campbell (1979), Simon (1978), and Laffont and Martimort (2001) suggest, continued to demonstrate increasingly distorted instructional practice would extend the thesis of this research and provide additional insight into the impact of accountability systems on instructional practice.

Implications for Practice

1. The implications for practice related to the findings of this study involve serious ethical challenges that educators face on a daily basis and will continue to face as long as student achievement on snapshot summative assessments is the metric by which educators are evaluated. Certainly, the literature cited in this study calls into question not only the accountability instrument itself, but also the validity of the tests used to apply the system as reliable measures of student achievement. Having been involved in the testing industry for many years, this is an especially personal issue for me. I have come to embrace more and more the concept of formative assessment, discussed in brief in this dissertation.

However, to truly be formative, testing cannot be punitive, as it naturally becomes whenever scores are disaggregated by school and teacher and reported publicly. When this happens, and principals and district administration naturally follow with programs to "improve" such scores, assessment ceases to be formative and devolves instead into a series of mini summative tests that differ from benchmarks and mock testing only in their length, frequency, and increased opportunities to feel pressured. The solution to this problem is obvious but

difficult. Though it is easy for theoreticians such as Stiggins (2005) to preach the perils of issuing grades and evaluating teacher effectiveness via such tests, it is far more difficult for school administrators to turn loose of this long time carrot and stick tool. Doing so will require a thorough understanding of the underlying deficiencies both in the accountability system and the assessments which inform it. As this study has attempted to show, however, there is more to the story of student achievement than whether or not a student can be counted in the proficient category for accountability purposes. In a society that values the individual and his or her independent pursuit of happiness, it is tragic that educators feel forced to pick and choose where to focus educational resources based on anything other than what is best for each individual child. It is my hope that this study and others like it, will encourage educational policymakers and practitioners alike to refocus our efforts not on numbers and rankings but on learning.

2. I have commented within this dissertation on several occasions about the unlikelihood that accountability systems will be going away anytime soon. With regards to the Jacob (2004) study there is certainly some question as to whether it should. Although the NCLB reauthorization has been stalled in Congress, its mandates continue to operate, other than in those few places that have so far applied for and been granted a waiver by the Obama administration pursuant to the president's recent executive order. Such waivers, however, come with their own strings and do not release educators from state level accountability mandates, most of which have continued to operate in a parallel fashion to NCLB.

To that end, if an accountability system is needed, as many including myself believe, it must be developed in such a way that it avoids many of the traps that we have discovered in the present systems. For example, it should not set a one-size fits all proficiency bar, but rather judge school efficacy on student growth. Likewise, such growth cannot be determined by a

single, annual multiple choice assessment that would then by its very nature become a high-stakes test, but should involve portfolio reviews of student work, including assignments, homework papers, classroom assessments, essays, projects, and other examples of student work that offer a truly holistic view of the progress a student makes over time, be the student a high or a low performer.

Such a system, of course, would be very expensive. However, like many accountability approaches, it would not necessarily have to be applied to all students every year. Schools would be responsible for keeping such a portfolio for all students, but accountability might be determined based on a random review of a small percentage of student portfolios. A second sample could easily be drawn from any schools that showed a pattern of problems in order to verify that any apparent discrepancies were systemic before imposing mandates for improvement.

Implications for Policymakers

The No Child Left Behind Act is today noted more for its intricacies than its idealism, but it should not be lost that the legislation marked the complete ascendancy of an educational philosophy that had been on the rise since the early years of the Reagan administration and by which United States policymakers with the support of the public put aside the input-based strategies and focus that had informed education policy for a quarter of a century in favor of outcome based education. As has been discussed herein, this was in many respects, a natural evolution that did encourage schools, many for the first time, to really pay attention to every student on the campus, and in this respect it is truly impossible to say the law had no positive effects.

However, charges that the law has engendered numerous and controversial unintended consequences, are equally difficult to dispute. This study sought to determine whether evidence existed that schools were engaging in distorted instructional practices as a result of trying to comply with the accountability mandates and how this practice might impact students. The results of the study thereby call into question whether NCLB is in fact helping to close the educational deficit long observed among disadvantage and limited English students or whether it is helping to mask continued problems behind a veil of impressive sounding numbers and statistics.

Although this study was limited in scope and precision, it highlights some of the underlying problems of using standardized tests as measures of school efficacy. Rothman (2004) and others have pointed to problems arising when educators try to employ state tests as synonymous with student learning. The underlying problems are only aggravated when the instrument exhibits deep flaws and establishes an unrealistic goal. Kim (2003) and others have pointed to NCLB's 100 percent proficiency threshold as an underlying cause of instructional distortion. The unintended consequences of policies, though well-meaning, typically disproportionately impact those individuals and institutions placed most at risk. Given that the accountability instrument arguably adds to the risk faced by these campuses, policy makers must pay close attention to the impact

of accountability policies on less privileged populations as well as the impact of policies designed to address systemic deficiencies deemed to be underlying causes.

Though Alexander (2003) as well as Reschovsky and Imazeki (2001) have rekindled the input- versus output-based policy debate, holding that lawmakers must do more to address inequities in the system and level the playing field for all students, I believe it is just as important

to acknowledge that while steps can be taken to mitigate the differences, such differences will always exist and should not be ignored or treated as nonexistent.

The underlying purpose of NCLB was to ensure that all students would have access to high quality schools, but regulatory policy clearly has not facilitated

that change. A decade after its implementation, the fact that government officials are predicting dramatic increase in numbers of schools failing to meet AYP is illustrative of the need for greater emphasis on the contextual differences perpetuated between schools and students due in part to the inequities built into the present system. Simply letting the legislation fade into the twilight and allowing it to take its place on the scrap heap of bad ideas, will do nothing to solve the issues that continue to plague public education. Policymakers must find ways to hold educators, students, and parents alike accountable in a way that is constructive for all, rather than for a limited cohort chosen consistent with a school's accountability risk. The goal of educating our children is far too important to accept anything less.

Concluding Remarks

As noted earlier in this dissertation, I have spent a good portion of my professional career involved in the development of standardized test items for state summative assessments. For more than a decade, I have worked with some of the largest test publishers in the country, as well as with state and local officials in several states to create the large scale assessments discussed at length in this dissertation. It has been an interesting journey to say the least. Therefore, as I noted earlier, the subject of testing and accountability and their impact on students is a deeply personal one for me. Obviously, I would not have pursued a career in testing, had I not believed strongly at the time that these tests and the accountability instruments which they inform were valuable tools for improving schools. Now, however, although I still believe that testing has its

place in the educator's tool box, and even that some form of accountability must be retained, I have over the past several years of personal observation, discussions with colleagues and professors in my doctoral program, and by talking to teachers, fellow administrators, students, and parents, come to realize that like the seemingly harmless Mogwals of Gremlins fame, accountability has devolved into a presence that has now come to overwhelm every thing else in education.

By the time I entered the doctoral program at U.T. El Paso, I was already beginning to question many of the underlying assumptions I had carried about accountability for many years. However, as evidence continued to mount that the charges by teachers of curriculum reduction, drill and kill mandates, and general testing chaos were in fact more than just the whiney complains of a few teachers trying to avoid the hard work necessary to educate their students, I was forced to begin reevaluating my beliefs. As I entered my second year of the doctoral program, I was holding onto the last rung of support for these policies, where I joined the chorus of other desperate holdouts singing "nobody forces them to teach to the test." While ostensibly this may be true, it does not alleviate the damage being done to our students by the application of this system, regardless of who may be to blame.

I was inspired to pursue this research in part because it took me so long to understand something that should have been obvious much sooner. As Ravitch and Finn (2007) have noted apologetically, "we really should have seen this coming" (p. 2). Nevertheless, I remain convinced that educational policy, as assuredly as it cannot be the panacea for all that ails education, can neither be the cause of all of its many problems. As educators and professionals, we are faced every day with choices that are profound in that their impact extends far beyond our own lives but have long-lasting impact on the lives of our students both in good ways and in bad.

Therefore, though it will be difficult, we must resist the natural instinct to make such decisions with our own interests placed above those of our students. It is easy to say we must, but we must. NCLB is not going away anytime soon. When it does, it will likely be replaced by another system. As with any system, said future system will no doubt have positives and negatives. Moving forward, whether working within our present system or someday within a new one, we must always keep the best interests of our students at the forefront. We must.

Chapter Summary

This chapter began with a description of the context of the study, then briefly reviewed the study questions and methods employed by the researcher. The chapter next discussed the theoretical framework as it relates to the study results and described some preliminary conclusions based on their analyses along with some suggestions for further related research. The chapter then reviewed the extant literature and its relationship to this study and its findings, before closing with a description of possible implications for practice and policy, followed with some final, concluding thoughts.

REFERENCES

Abedi, J. (2004). The No Child Left Behind Act and English Language Learners: Assessment and Accountability Issues, *Educational Researcher*, Vol. 33, No. 1.

Alexander, N. A. (2003). Considering equity and adequacy: An examination of the distribution of student class time as an educational resource in New York state, 1975-1995. *Journal of Education Finance*, 28(3), 357-381.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Ballou, D., & Springer, M. (2008). Achievement trade-offs and No Child Left Behind. *National Center on School Choice*, Vanderbilt University, Nashville, TN

Amrein A. L., & Berliner, D., C. (2002). High-stakes testing, uncertainty, and student learning, *Education Policy Analysis Archives, 10* (18). Retrieved October 4, 2011 from http://epaa.asu.edu/epaa/vl0nl8/.

Amrein A. L., & Berliner, D., C. (2002). An Analysis of Some unintended and negative consequences of high-stakes testing. Great Lakes Center for Educational Research and Practice. Retrieved October 6, 2011 from http://www.greatlakescenter.org/pub/H-S%20Analysis%20final.pdf

Bailey, A.L., & Butler, F. A. (2003). An evidentiary framework for operationlizing academic language for broad application to K-12 education: A design document (CSE Tech. rep. 611) Los Angeles, University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Bishop, J., & Mane, F. (2001). The impacts of minimum competency exam graduation requirements on college attendance and early labor market success of disadvantaged students. In G. Orfield &M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 51-84). New York: Century Foundation.

Bodgan, R.C., & Bilklen, S.N. (2003). *Qualitative research for education: An introduction to theories and methods (4th ed.).* New York: Allyn & Bacon.

Bratt, L., Kim, J., & Sunderman, K. (2005). English language learners: Increased accountability under NCLB. Retrieved April 26, 2009 from http://www.civilrightsproject.harvard.edu/research/esea/LEP_Policy_Brief.pdf

Braun, H. (2004). Reconsidering the impact of high-stakes testing, *Education Policy Analysis Archives, 12(1).* Retrieved October 4, 2011 from http://epaa.asu.edu/epaa/vl2nl/

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning* 2: 67-90.

Carnoy, M. (2005). Have state accountability and high-stakes tests influenced student progression rates in high school? *Educational Measurement: Issues and Practice, 24(4),* 19-31.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24(4),* 305-331.

Cohen, D.K., & Moffitt, S. L. (2009). *The ordeal of equality: Did federal regulation fix the schools?* Cambridge MA: Harvard University Press.

Crocker, L. (2005). Teaching for the test: How and why test preparation is appropriate. In Phelps, Richard P. (Ed.) *Defending standardized testing*. Mahwah, NJ, Lawrence Erlbaum Associates.

Cronin, J. Dahlin, M., Xiang, Y., & McCahon, Donna (2009). *The accountability illusion*, Thomas B. Fordham Institute, Washington D.C.

Cummins, J. (1995). Empowering minority students: A framework for intervention. In O. Garcia & J. Baker (Eds.) *Policy and practice in bilingual education: Extending the foundations* (pp. 10-116). Clevedon, English: Multilingual Matters.

deMarrais, K., & Laplan, S.D. (2004). Introduction. In K. deMarrais & S.D. Laplan (Eds.), *Foundations for research: Methods of inquiry in education and the social sciences* (pp. 1-12). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers

Dee, T., & Jacob, B. (2010). The impact of No Child Left Behind on student achievement *Journal of Policy Analysis and Management*, Vol. 30, No. 3, 418-46.

Donovan, C., Figlio, D., & Rush, M. (2007), Cramming: The effects of school accountability on college-bound students. Washington, D.C.: *National Center for Analysis of Longitudinal Data in Education Research*, Urban Institute

Dow, Gregory K. (2003). *Governing the firm: Workers' control in theory and practice.* Cambridge UP

Editorial Projects in Education Resource Center. Quality Counts National Highlights Supplement. (2008). *Education Week*, url: http://www.pewtrusts.org /uploadedFiles/wwwpewtrustsorg/Reports/K-12_education/National %20Highlights%20Report.pdf on January 10, 2012.

Erpenbach, W., Forte-Fast, E., & Potts, A. (2003). Statewide educational accountability under NCLB: Central issues arising from an examination of state accountability workbooks and U.S. Department of Education reviews under the No Child Left Behind Act of 2001, State Collaborative on Assessment and Student Standards, Council of Chief State School Officers (CCSSO).

Farr, B. & Trumbull, E. (1997). *Assessment alternatives for diverse classrooms*. Norwood, MA: Christopher Gordon Publishers.

Field, A. (2009). *Discovering statistic using SPSS **(3rd** ed.).* London: SAGE Publications Ltd.

Finn, C. E., Jr., & Ravitch, D. (2007). *Beyond the basics: Achieving a liberal education for all children.* Thomas B. Fordham Institute.

Freeman D. & Freeman, Y. (1998). *ESL/EFL Teaching: Principles for Success*. Portsmouth NH.: Heinemann.

Gillespie, R. (1991). *Manufacturing knowledge: A history of the Hawthorne experiments*. Cambridge UP

Guggino, P. & Brint, S. (2010). Does the No Child Left Behind Act help or hinder K-12 education?" *Policy Matters*, Vol. 3, No. 3, California Riverside UP.

Hanushek, E. A. & Raymond, M. E. (2003a). Improving educational quality: How best to evaluate our schools? In Y. Kodrzycki (Eds.), *Education in the 21st century: Meeting the challenges of a changing world.* Federal Reserve Bank of Boston.

Hanushek, E. A. & Raymond, M. E. (2003b). Lessons about the design of state accountability systems. In *No Child Left Behind? The Politics and Practice of Accountability*,edited by Paul E Peterson and Martin R. West. Brookings.

Hanushek, E. A. & Raymond, M. E. (2004). The effect of school accountability systems on the level and distribution of student achievement. *Journal of the European Economic Association,* 2(2-3), 406-415.

Hoffer, T., Hedberg, E.C., Brown, K, Halverson, M.L. & Reid-Brossard, P. (2011). F*inal report on the evaluation of the growth model pilot project,* National Opinion Research Center, University of Chicago, U.S. Department of Education

Holmstrom, B., & Milgrom, P. (1994). The firm as an incentive system. *The American Economic Review*. Vol. 84, No. 4. url: http://www.jstor.org/ stable/2118041

Jacob, B. A., (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools." *Journal of Public Economics* 89 (5-6), June:761-796.

Jones, G. J. (1995). *Organizational theory: Text and cases.* New York, Addison-Wesley Publishing Company

Johnson, B. (2001). Toward a new classification of nonexperiemental quantitative research. *Educational Researcher*. (30)2,3 - 13.

Kane, T.J., & Staiger, D.O. (2001). Volatitlity in school test scores: Implications for test-based accountability systems. *Brookings Papers on Education Policy*, 5, 235-280.

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs: analysis of variance and multiple regression/correlation approaches.* New York: W.H. Freeman & Company

Kim, J. (2003). *The initial response to the accountability requirements in the No Child Left Behind Act: A case study of Virginia and Georgia.* Unpublished manuscript.

Kim, J. & Sunderman, G., (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher*, Vol. 34, No. 8, pp. 3–13

Koretz, D.M. (2000). Limitations in the use of achievement tests as measures of educators' productivity, *RAND Education Center for Research on Evaluation, Standards, and Student Testing*

Laffont, Jean-Jacques and Martimort, David. (2001). *The Theory of Incentives: The principal-agent model.* Princeton, NJ: Princeton UP.

Lessow-Hurley, J. (2003). *Meeting the needs of second language learners: An educator's guide.* Alexandria, VA.: Association for Supervision and Curriculum Development.

Linn, R. L. (2000). Assessments and accountability. *Education Researcher, 29(2),* 4—15.

Linn, R.L. (2009). The concept of validity in the context of NCLB, in *The concept of validity: Revisions, new directions, and applications,* ed. Robert W. Lissitz, Information Age Publishing, Charlotte, NC

McLaughlin, D, Mello, V., Blankenship, C, Chaney, K., Esra, P., Hikawa, H., Rojas, D., William, P., & Wolman, M. (2008). *Comparison between NAEP and State Reading Assessment Results: 2003* (NCES Rep. No. 2008-474).Washington, D.C.: National Center for Education Statistics.

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. Peabody Journal of Education, 65(3), 29-46.

McNeil, L., & Valenzuela, A. (2000). *The harmful impact of the TAAS system of testing in Texas.* Cambridge, MA: Civil Rights Project at Harvard University.

McNeil, L. (2005). Faking equity: high stakes testing and the education of Latino youth. *Leaving children behind,* Angela Valenzuela, ed., State University of New York Press

McNeil, L. (2000). *Contradictions of school reform: Educational costs of standardized testing.* New York: Routledge.

Nichols, S. L., Glass, G. V, & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education*

Oakes, Jeannie. (2005). *Keeping track: How schools structure inequality*. (2nd edition). New Haven, CT: Yale University Press.

Padilla, R. V. (2005). High-stakes testing and educational accountability as social constructions across cultures. In A. Valenzuela (ed.) *Leaving Children Behind,* New York: State University of New York Press

Pearl, Judea, (2009). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press

Peterson, P. E., & Hess, F. M. (2006). Keeping an eye on state standards. A race to the bottom? *Education Next, 3,* 28-29.

Policy Analysis Archives, 14(1). Retrieved February 15, 2012 from http://epaa.asu.edu/ epaa/v14n1/.

Padilla, R. V. (2005). High-stakes testing and educational accountability as social constructions across cultures. *Leaving Children Behind*, Angela Valenzuela, ed., State University of New York Press

Quality Counts National Highlights Supplement 2008 (2008). Education Week, Editorial Projects in Education Resource Center. Retrieved from: http://www.pewtrusts.org/uploadedFiles/wwwpewtrustsorg/Reports/K-12_education/National%20Highlights%20Report.pdf on January 31, 2010.

Raymond, M.E., & Hanushek, E.A. (2003). High-stakes research. *Education Next* 5(3), 48-55.

Ravitch, D. (1996). *National standards in American education: A citizen's guide*. Washington,DC: Brookings Institution.

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

Rebell, M. and J. Wolff. (2009) *Moving Every Child Ahead*, Teacher's College Press, New York

Reschovsky, A., & Imazeki, J. (2001). Achieving educational adequacy through school finance reform. *Journal of Education Finance*, 26(4), 373-396.

Reyhner, J. and Singh, N. (2010). Aligning Language Education Policies to International Human Rights Standards. eJournal of Education Policy, Northern Arizona University. Retrieved December 12, 2011 from https://www4.nau.edu/cee/jep/journals.aspx?id=324

Rippberger, S. J. and Staudt, K. A. (2003). *Pledging allegiance: Learning Nationalism at the El Paso-Juarez border*. RoutledgeFalmer, New York

Rothman, R. (2004). Benchmarking and alignment of state standards and assessments. In S. H. Furman & R. F. Elmore (Eds.), *Redesigning accountability systems for education.* New York: Teachers College Press.

Rothstein, Richard. (2008). Holding accountability to account: How scholarship and experience in other fields inform exploration of performance incentives in education. *National Center on Performance Incentives*. Vanderbilt Peabody College.

Rosenshine, B. (2003, August 4). High-stakes testing: Another analysis. *Education Policy Analysis Archives*, 11(24). Retrieved January 7, 2012 from http://www.aspeninstitute.org/policy-work/no-child-left-behind/reports/state-standards-assessing-differences-qual

Saminsky, Alina. (2011). The Reauthorization of No Child Left Behind, and Avenues for Improvement. *Student Pulse*, 3.01. Retrieved November 21, 2011 from http://www.studentpulse.com/a?id=375

Simmons,W., & Resnick, L. (1993). Assessment as the catalyst of school reform. *Educational Leadership*, February, 11-15.

Simon, Herbert A. (1978). "Rational Decision-Making in Business Organizations." Nobel Memorial Lecture. December 8. url: http://nobelprize.org/nobel_prizes/economics/laureates/1978/simon-lecture.pdf

Spalding, E. (2000). Performance assessment and the new standards project: A story of serendipitous success. *Phi Delta Kappan*, *81*(10), 758-764.

Stecher, Brian and Sheila Nataraj Kirby. (2004). "Organizatonal improvement and accountability: Lessons for education from other sectors." William and Flora Hewlett Foundation. Rand Education, Santa Monica, CA

Stecher, Brian and Hamilton, Laura S. (2002). Putting theory to the test: Systems of accountability should be held accountable. *Rand Review* Vol 26, No. 1 Santa Monica, CA

Stiggins, Rick. (2005). From formative assessment to assessment for learning: A path to success in standards-based schools. *Phi Delta Kappan* Vol 87, No. 4,

Stone, Deborah. (2002). *Policy paradox: The art of political decision making*. New York: Norton & Company.

Stotsky, S. (2000). *What's at stake in the K-12 standards war*. New York: Peter Lang.

Suskie, L.A. (1996). Questionnaire survey research that works. Tallahassee, Fl.: Association for Institutional Research.

School Data Direct (2009). Council of Chief State School Officers through Standard and Poors, url: http://www.schooldatadirect.org/ on April 25, 2011

School Enrollment 2009 (2009). United States Census Bureau publications.

Thomas, W. P. & Collier, V. P. (1997). *School effectiveness for language minority students.* Washington, DC: National Clearinghouse for Bilingual Education

Teachers of English to Speakers of Other Languages (2003). *Position Paper on high-stakes testing for K-12 English-language learners in the United States of America*. Retrieved April 25, 2011 from http://www.tesol.org/s_tesol/bin.asp?CID=32&DID=375&DOC=FILE.PDF

Texas Administrative Code, Texas Secretary of State. url: http://www.sos.state.tx.us/tac/ on April 29, 2011

Texas Education Agency (2011). The 2011 Accountability Rating System for Texas Public Schools and School Districts, Texas Education Agency Department of Assessment, Accountability, and Data Quality Division of Performance Reporting

Texas Education Agency (2011). 2011 Adequate Yearly Progress (AYP) Guide for Texas Public Schools and School Districts, Texas Education Agency Department of Assessment, Accountability, and Data Quality Division of Performance Reporting *Texas study of students at risk: Case studies supporting ninth graders' success*. (2004) Texas Education Agency

The Education Trust (2011). *Getting it right: Crafting federal accountability for higher student performance and a stronger America*. Washington, D.C.

Tracey, C., Sunderman, G. & Orfield, G. (2005) C*hanging NCLB district accountability standards: Implications for racial equity.* Cambridge, MA: The Civil Rights Project at Harvard University

US/Mexico Border Counties Coalition. (2006) *At the Cross Roads: US / Mexico Border Counties in Transition*

Viadero, D. (1994, July 13). Teaching to the test. *Education Week*, XIII(39), 21-25.

Wilson, L. W. (2002). *Better instruction through assessment*. Larchmont, NY: Eye on Education.

Wright, W. (2002). The effects of high-stakes testing in an inner-city elementary school: The curriculum, the teachers and the English language learners. *Current Issues in Education* 5 (5). Retrieved April 23, 2011 from http://cie.ed.asu.edu/volume5/number5/

CURRICULUM VITA

Curtis Barnes was born in San Diego, California to Jack and Rilla Barnes. He graduated from Andress High School in El Paso, Texas in 1980. He enrolled at the University of Texas at El Paso (UTEP) in 1985. Curtis majored in journalism, earning a Bachelor's from the College of Liberal Arts in 1993. He earned a Master of Fine Arts from UTEP in 1997 and a Master of Education from UTEP in 2000. He entered the doctoral program in Educational Leadership at UTEP in 2008. Curtis began teaching English and reading with the El Paso Independent School District (EPISD) in 1995 and at the time of this research worked in the Research and Evaluation department of the EPISD.

Permanent Address: 10804 Loma de Alma, Dr.
El Paso, TX. 79934