" . . . BUT THE LIGHT'S BETTER HERE," HE SAID: MISCHANCE, MISAPPLICATION AND MISDIRECTION IN DATA-DRIVEN APPLICATIONS

by

RAMONA LOWE

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2012

ACKNOWLEDGEMENTS

A funny thing happened on the way to this dissertation: life. I'd like to thank everyone who supported me during all of those life issues (and there were some big ones) that popped up in the years it has taken to reach this place. A super-sized thank you is due my friend, writing partner, and fellow cohort member, Dr. Sarah Fitzhugh. I truly would not have made it without you. I'd also like to give a deep and heartfelt thanks to my advisor, Dr. Jeanne Gerlach, whose support and encouragement made this process so much more meaningful, not to mention manageable. Her encouragement was a constant. I'd also like to thank Dr. Jim Hardy and Dr. Holly Hungerford-Kresser who served on my committee for their help and guidance. Thanks also to the administration and teachers at Lewisville ISD. Thanks are especially due to my supervisor and friend, Donna Henry, and my lifesaver, Courtney Hart. My family has also been supportive, albeit across the miles. I know my mom would have been proud had she lived to see this day. And of course, there are the odd acknowledgements. Thanks to the fine folks at Panera Bread #1385 for providing such a great atmosphere for writing and fantastic chicken noodle soup. A shout out to the roomiest Starbucks around on Grapevine Mills Parkway. A Frappucino always helped. Three visits to Estes Park, Colorado, helped a great deal, and I grant Estes the title of best place in the world to "dissertate" award.

Finally, for Rigoberto Ruelas, who reminds us all that this is for real.

July 16, 2012

ABSTRACT

" . . . BUT THE LIGHT'S BETTER HERE," HE SAID: MISCHANCE,

MISAPPLICATION AND MISDIRECTION IN

DATA-DRIVEN APPLICATIONS

Ramona Lowe, PhD

The University of Texas at Arlington, 2012

Supervising Professor: Jeanne Gerlach

The primary purpose of this study was to provide quantitative data on the use of released Texas Assessment of Knowledge and Skills (TAKS) exams as benchmark test instruments in a sample population. This study was specifically undertaken to determine the extent of this practice, especially in light of published material from Texas Education Agency (TEA) that states this practice is not an appropriate use of the released instrument. An additional purpose was to gain quantitative data on the use of data harvested from both benchmark and actual administrations of TAKS instruments in decision-making. District level supervisors/coordinators of English/language arts from a geographic target area were chosen to be the sample population for this

survey questioning the use of released TAKS exams and the subsequent use of data harvested from those benchmark assessments, as well as the use of actual TAKS scores. Survey questions were developed and tested for validity and reliability before being sent to the survey pool of 75 educators from districts in a target area (combined student population of 625, 797). The response rate was 52%. The percentage of respondents who indicated that their districts used released TAKS tests as benchmark instruments was 68.3%, with 80.5% disaggregating the data from both these administrations and actual TAKS tests. That number rose to 98.6% when adding those who said they only disaggregated data occasionally. Only 1.0% of respondents said they never disaggregated TAKS data. However, not quite half (47.5%) said their districts always or frequently used TAKS data alone to make important decisions regarding student interventions. Slightly more than one-third (37.5%) indicated that the results of a benchmark exam had an impact on district instructional/curricular decisions. Over half (52%) indicated their districts used the released TAKS exams in full-scale rehearsals that significantly interrupted the school day to prepare for actual administration.

The implication is more districts are using multiple data sources to inform their data-driven applications, even though many districts utilize released TAKS exams as benchmark instruments. Further research is suggested regarding the assessment literacy of educators, particularly in relation to the use of a summative test for harvesting diagnostic data as Texas transitions to the new STAAR/EOC exams.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

CHAPTER 1

INTRODUCTION

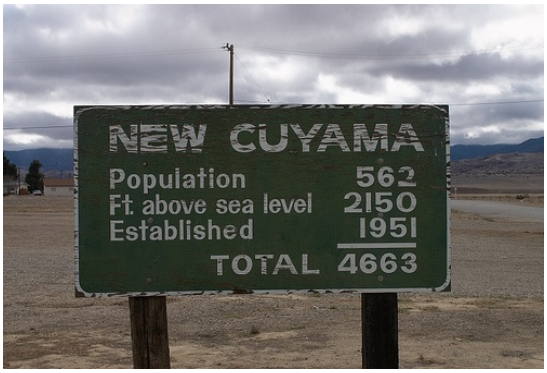

Figure 1.1 The Math Is Correct, But Do These Numbers Add Up?
Photo Credit Josh Hill, Flickr.com via Creative Commons.

1.1 "Data, Data, Everywhere . . ."

When *Educational Leadership* devoted an entire issue to data in the winter of 2008, W. James Popham, respected researcher, author and Professor Emeritus at UCLA, referenced Coleridge's lines from "The Rime of the Ancient Mariner," but construed them to match the frenzy over test data: "Data, data, everywhere/Too much to let us think" (p. 85). It is a valid concern in the post-No Child Left Behind (NCLB) landscape of American K-12 education. Ushered to supremacy by the historic legislation in 2002, data has been given unprecedented power to determine success, failure, and—in some cases—even existence of schools. Strongly tied to the so-called "business model" that espouses that schools should be run as businesses, every facet of decision making, from curriculum to discipline to assessment to the length of the school

day, was to be made logically based on data tied to student achievement (Koretz, 2008; Ravitch, 2010). School systems were now considered "data-driven" with the thinking that judgments based on data—hard, unbiased, scientific data, gathered almost exclusively from the high-stakes tests NCLB demanded that each state create—would produce optimal results: increased student achievement. Data could quantify and clarify complex situations and unequivocally point the way to the appropriate corrective actions. The stunning simplicity and obviousness of the concept makes one wonder why it was not thought of before.

1.1.1 Too Much to Let Us Think

When considering the data-driven approach in education, it would be prudent to take H. L. Mencken's oft-quoted witticism on simple solutions into account: "For every complex problem, there is an answer that is clear, simple, and wrong" (as cited in Burke, 2002). In addition, Koretz's (2008) opening comments in *Measuring Up: What Educational Testing Really Tells Us* (2008), convey that this "apparent simplicity, however, is misleading. Achievement testing is a very complex enterprise, and as a result test scores are widely misunderstood and misused" (p. 1). Ravitch, an educational historian and former member of the first President Bush's Department of Education, had an abrupt about face on the testing and charter school aspects of NCLB. Ravitch (2010) writes in *The Death and Life of the Great American School System* that she "now found [her]self experiencing profound doubts about these . . . ideas" (p. 1). Lest anyone

doubt the sincerity of her reversal from her once-staunch support for NCLB (she was present at the law's signing), she writes just a few pages later that,

> I was concerned that accountability, now a shibboleth that everyone applauds had become mechanistic and antithetical to good education. Testing, I realized with dismay, had become a central preoccupation in the schools and was not just a measure, but an end in itself. (Ravitch, 2010, p. 12)

Ravitch goes straight to the heart of the matter when she writes that she began to look at "schools and teachers and students from an altitude of 20,000 feet and seeing them as objects to be moved around by big ideas and plans" (p. 10). In short, Ravitch was looking at schools like those in federal and state governments do, like those in school reform movements do, and like those who stand to make money off school reform movements do. Ravitch was approaching the matter as one who consults statistical models in place of direct contact with actual students, teachers, and situations. Far removed from the daily realities of actual schools, their vantage point distorts reality.

Mencken's (as cited in Burke, 2002) statement about simple but wrong solutions for complex problems is even more apropos when looking at the large body of work on the subject from W. James Popham, who bluntly states that data tell us nothing: it is the *inferences* drawn from the data that drive decisions. For all that schools claim to be—data-driven, no such animal exists; it is the inferences drawn from that meaningless-in-itself data that fuel decisions. But what guarantees that those inferences are sound? What requires that those inferences represent a logical, and more importantly, an

accurate interpretation of the data? What assures that those inferences do not represent an oversimplification of a multi-faceted educational program? In a word, *nothing.* The status quo does not have measures in place to catch faulty uses of data.

Popham (2009) asserts that, "most of today's educators know almost nothing about educational assessment" (p. 126). Popham explains that, while educational programs are focused on other concerns (curriculum, discipline, finances, etc.), the biggest reason for this lack of assessment understanding is that, "many candidates for teacher or administrative credentials are intimidated by any content thought to be even mildly mathematical in nature" (p. 127). Koretz (2009) discovered much the same thing in his class in Harvard's graduate programs for school administrators. His students did not have what Popham (2008) calls "assessment literacy" and were using the powerful tool of data without any sense of what those numbers did and did not mean. In the pre-NCLB days, this lack of "assessment literacy" was not much of an issue, but in today's educational climate, the misunderstanding and/or misrepresentation of data exacts a high toll.

1.1.2 What is the Harm?

Education is too complex, too situational, and too affective for simple solutions. When we try to impose those simple solutions, the results are often worse than the problems they were meant to solve. The research of Berliner and Nichols (2007) explored the impact of high-stakes testing. They wrote that NCLB was "designed to push lazy teachers and lazy students to work harder" (p. 12). Its design was inherently

punitive. These researchers, Berliner especially, have been studying the effects of high-stakes testing since before NCLB and have concluded that the atmosphere spawned by that seemingly logical, simple notion of basing decisions on data from these tests is not only ineffective, it is toxic. The title of Berliner and Nichols' book, *Collateral Damages: How High-Stakes Testing Corrupts American Schools*, lays out their findings, which include an epidemic application of Campbell's Law (more on this shortly) and a negative impact on student learning.

In her book, *Wounded by School*, Johnson (2009) wrote of the wholesale harm of standardized testing upon students, stating that it was not just the testing itself that brought harm—though there were plenty of stories of individual harm and harsh condemnation for the tests themselves:

> [T]hey were also wounded by the sense that our educational system learning is regarded as a product, not a journey; that every learning experience can be assessed in a way that produces easy-to-report data; and that they themselves were unformed products ready to be shaped and stamped and graded within the institution. (Johnson, 2009, p. 28)

Advocates of high-stakes testing will, when pressed, accept that not everything that is important in education can be measured with a quantitative assessment, but then declare that we must work with what we have and that is quantitative data. That idea fits with Ravitch's 20,000 feet above metaphor: quantitative data is easily visible and theoretically comparable from school to school. However, that view is distorted, and

that argument is too dismissive of the aspects of learning that cannot be measured quantitatively. A school is far more than the sum of a set of numbers. Nichols and Berliner (2007) countered that type of number-oriented thinking with Einstein's quote: "Not everything that counts can be counted, and not everything that can be counted counts" (p. 46). Furthermore, quantitative data, as will be seen, is not the objective measure that it is presented to be. Chapter 2 details the limitations of quantitative data, which for this study were defined as test scores from mandated testing.

It is not my intent to claim that these test scores are without value or that data do not belong in the current educational configuration. It can, if used properly, be an illuminating instrument that helps all parties create a better education for American students. However, data should not drive, but inform. The term "data-driven" should be accompanied by a critical warning that is often overlooked. Chapter 2 covers this concept in more depth.

For all the talk about data being used to drive decision making, data is, ironically, a nebulous term. By definition it is plural, but when multiple data sources are brought to the table, a clear measure is not always possible. Enter high-stakes tests, which in Texas has meant Texas Assessment of Academic Skills (TAAS), then Texas Assessment of Knowledge and Skills (TAKS), and, shortly, State of Texas Assessment of Academic Readiness (STAAR) including End of Course exams (EOCs). These tests carry a disproportionate weight in shaping the education of children, because those scores fit the parameters of the data-driven concept perfectly—in theory. The tests are

standardized but not normed, an important distinction. TAKS is a criterion-referenced test tied to the state standards (the TEKS), whereas norm-reference tests derive scores from a comparison group and are rankings. From these results, campus and district accountability is determined. Among the power players--that is, policy makers, the media, those driving educational reform--no one asks about the soundness of inferences from these scores generated by campuses, districts, and even the state. No one asks if the data is being used in ways that are in accord with the test design. No one asks if these disaggregated and drilled-down pieces of data are valid to inform instruction, and if so, how? No one asks if there are more precise sources of data. To borrow a mariner's metaphor, no one asks about the sea-worthiness of these vessels. But they should, and if they would, they'd hear those questions being asked by those who live closest to the tests: students, parents, teachers, and campus administrators who live with the fallout.

Popham and others have written that no one test can do everything, and yet, TAKS is the assessment version of one-stop shopping. By state law, it has been used for promotion in Grades 3, 5, 8 and 11 (exit level). It is used to determine if a campus or district is exemplary, recognized, acceptable, or unacceptable. It is also used to place individual students in remedial classes or other such interventions that limit student electives. Tovani, a teacher-author from Colorado, reported last school year that her students were "angry" on the first day of school when they arrived at her high school reading class and justifiably so: they had not had an elective course in three years because they had not passed the test. Their elective classes were remedial math, reading,

and test prep. (C. Tovani, personal communication, October 15, 2009). The simple solution "fail the test = remedial class" had failed to work, and had restricted students from elective courses that might actually have helped them improve both their reading skills and their attitudes toward school. The punitive nature of the status quo is evident. There are such stories from all around the country, and especially from Texas where an unholy attachment to TAKS scores is at the root of any number of school decisions, including which electives (or recess periods in elementary) a student is permitted to take. Well-meaning principals and counselors feel, as one principal I have worked with said, "When the choice is between learning to read and PE, I'll take reading." The student then has two reading classes (the required by the state and the extra from the school) and while that sounds good (double time for reading), the truth is that middle school children need what PE offers. A school day with eight periods of traditional classes is a trial for any student, and is excruciating for the struggling student.

K-12 schools are making drastic efforts to do "whatever it takes" to achieve higher scores. One at-risk elementary school in my district has reduced recess to once a week because the time "cannot be spared from instruction." (J. Pribanic, personal communication, May 2010). While TAKS scores hit record heights for the 2009-2010 school year (TEA 2010); the number of students requiring remedial coursework at the college level has also increased (Hacker, 2010). Those high TAKS scores were ultimately revealed to be a statistical sleight-of-hand (*Houston Chronicle*, 2010), but the

numbers needing remedial work in college remain as a testament to the failure of high-stakes testing to do what it was meant to do: guarantee student achievement.

Consider also the area of writing. Writing well upon demand is a critical skill for success in college and the workplace. K-12 education has the mandate of preparing students for that high level of writing, but the obsession with TAKS scores has severely limited students' ability to write. Take, for example, the writing tests that are part of TAKS in Grades 4, 7, 10, and 11 (exit level). Since 2003, the composition mode for all levels has been a narrative because that is what is written into the test design. Since it cannot be anything else (expository, persuasive, analytical, etc.), other writing is not privileged in the classroom. Students get endless practice in writing narratives, or rather, one narrative. As TAKS season approaches, teachers begin prepping the students by having them select one event from their lives that can be used with any prompt and to twist their narrative to fit the prompt. For example, an afternoon of fishing with grandfather can be used for the prompts: "Describe a favorite memory," "Write about a time you spent with someone you respect," "Write about a time you learned an important lesson," "Discuss the importance of having a hobby," "Write about your favorite place," etc. Since the composition has been practiced so many times before, students are also well rehearsed in what figurative language to use to describe people and places, how to "explode the moment" by describing what is seen, heard, felt, tasted, and so on. There is little true spontaneity or writing skill expressed. Why not? It is simple: teachers do not encourage students to take risks here because the stakes are too

high. This test is not about showing the writing abilities of Texas students, it is about gaming the situation. Too much is riding on the outcome of the test.

We have figured out how to beat the test, and do beat it consistently. Writing has one of the highest passing rates of all the TAKS subjects. Are those Texas students truly proficient writers? Ask freshman composition teachers at Texas colleges or even high school teachers. They would not agree. (The future STAAR/EOCs call for an expansion to additional genres, tilting the board so to speak. It will be interesting to see how this change will play out.) This area is more deeply explored in Chapter 2.

I write these things not to criticize teachers or Texas schools. They are playing the hand they have been dealt. The pressure to deliver is intense and ever increasing. I personally have used the twist-the story strategy, and died a little inside each time because it and other test-successful strategies are anathema to good teaching and student need. Those experiences, in part, are what drove me to this research. No one was asking if what it takes to get these high scores is best for students and for campus/district writing programs. In theory, students who have abundant writing experiences across the years should do well on this test, but the reality is that schools and teachers have discovered that the narrow focus works well to deliver high-test scores and are either unwilling or afraid to deviate from that model. The axiom "what is tested is taught" is proved to be true here with potentially crippling limitations for students.

1.2 Statement of the Problem

Overnight, it seems, school administrators were told to be data-driven and to dig deeply into test scores to discover the success or lack of success of their efforts. On the surface, it seems logical, but educational measurement is not easily accomplished or interpreted, and many, if not most, administrators do not truly understand educational data and the limitations on its use (Koretz, 2009). Due to a lack of what Popham (2008) calls "assessment literacy" (p. 125), data from TAKS is being used in ways that are not in alignment with the test design, causing harm to the education of Texas students. One such use is score data produced from TAKS being used as the sole criteria for important decisions. Another is the use of released TAKS tests as diagnostic instruments/ benchmarks. Both of these practices are highly questionable and have only limited practical use, yet they are used as the basis for educational decisions that are, in Nichols and Berliner's (2007) term, "corrupting" Texas education. This study seeks to understand how widespread these two practices—using released TAKS tests as benchmarks and using TAKS data as the data source for making decisions—are in a representative sample of districts in the North Texas area.

1.3 Research Questions

I adopted the viewpoints, briefly outlined in this introduction and discussed completely in Chapter 2, that the data-driven model was counterproductive, and that this was a major issue/problem in Texas education. The research question, then, was how widespread are these practices? While the greater issues of data-driven manifestations

informed and guided this study, I intended to focus on three questions in my research. The first was how many school districts in the sample group were using TAKS data alone (either from the actual administration or from a released test given as a benchmark or combination of the two) to make important decisions such as placing students or evaluating instruction. The second question was how widespread was the practice of using a released TAKS as a diagnostic benchmark. The third question dealt with how survey participants felt about the effects of TAKS and concerns for STAAR.

The first question addressed the role of TAKS scores on campuses. TEA's own technical manuals on TAKS (2010) call for multiple sources of data before making any significant decision regarding either a child or a program, yet the reliance on a single source (TAKS score) was perceived to be endemic. This study attempted to provide some quantitative data on just how prevalent this practice was in the sample.

Whenever teachers congregate, in both the physical and cyber worlds, there is complaint about excessive testing of students in preparation for TAKS. One of the most frequent complaints lodged is against "benchmarking," which is often the use of a released TAKS test to generate prescriptive data. It seems very logical to use TAKS in this pretest/posttest fashion, but once again, this simple response to a complex situation is the wrong one. There are fundamental differences between formative and summative assessments. The formative assessment is designed to provide information to guide instruction and will have multiple questions on just a few concepts. The summative assessment will have fewer questions per concept (sometimes as few as only one), but

assess more concepts—exactly like TAKS. Popham's (2009) argument about "abysmal assessment literacy" was found to apply here. It was a misuse of the released TAKS test to use it as a diagnostic benchmark (Carnegie Council, 2010; TEA, 2010). This study again sought to find how prevalent this practice was in schools in Texas.

1.4 Definition of Terms

Within this document, I am defining the following terms thusly. While others may use slightly different wording to describe these organizations and concepts, these definitions provide the reader with my conceptions of these items.

Accountability Rating: In Texas, campuses and districts are issued a rating based on their TAKS scores and dropout statistics. The ratings are as follows: exemplary, recognized, acceptable, and unacceptable. The rating levels are achieved by having all population groups on a campus meet the minimum passing standard. In 2009-2010, these ranged from 55% to 70% for acceptable. To achieve exemplary, schools needed to have 90% of all population groups meet the standard.

Data: In this document, data most often relates to the information generated in numerical form from the high-stakes tests mandated for each state by NCLB, primarily TAKS. As such, it was considered a mass noun and therefore referenced as a singular noun. The exception is when data from multiple, conflicting sources is referenced, as in "When Data Collide."

Diagnostic Assessment: Also known as formative assessment or benchmarking. Its purpose was to diagnose specific student abilities and learning so that instruction/

intervention can be better planned. This was preliminary assessment; more assessment followed as instruction proceeded. It is important to note that the focus of this research did not deal with the worth or validity of the concept of a diagnostic test or benchmark, the concern was with the improper (according to test design) use of released TAKS tests as instruments for benchmarks or other diagnostic assessments.

Important decisions: For the purpose of this study, important decisions are defined as those (a) affecting student placement in intervention efforts (remedial coursework, reading labs, tutoring programs, test prep programs, etc.); (b) affecting teacher evaluation (effectiveness, duty assignment, special pay, program evaluation, etc.); and (c) affecting curriculum content and evaluation.

No Child Left Behind (NCLB): Landmark bipartisan legislation signed into law in January 2002. Among the features is increased accountability for schools through testing (standards and tests developed by each state).

Professional Organization: My target population was a regional affiliate of a professional organization comprised of English/language arts supervisors/coordinators. The anonymity of this group will be maintained by this designation.

Summative assessment: Its purpose in this study was to evaluate learning after the fact, and it was a terminal assessment such as a final exam, end of course exam, or the tests required by NCLB. The intention was to see what students *had* learned, or *were not* learning. The use of a summative instrument for a diagnostic purpose was one of the major areas of this study.

Texas Assessment of Knowledge and Skills (TAKS): Standardized test used in Texas to replace Texas Assessment of Academic Skills (TAAS) since 2003. It serves as the accountability feature demanded by NCLB, but will be phased out by a new testing program (STAAR) beginning in 2012 (TEA, 2010).

Texas Education Agency (TEA): Office of the state government that oversees public education in the State of Texas.

Texas Essential Knowledge and Skills (TEKS): The curriculum for the state of Texas. Curriculum for each subject is divided into an opening statement, a knowledge and skill statement, and then student expectation statements (SEs). On TAKS scoring reports, the SEs are grouped into objectives.

1.5 Purpose of the Study

The purpose of this study was to survey the English/language arts supervisors from The professional organization and surrounding areas to find out how both released TAKS reading tests and reading test scores were used in their districts. These two areas were, I believe, at the root of two of the most egregious misuses of data: the use of a summative test to produce diagnostic data and the use of a single score as the basis for important decisions on a school campus. Neither was an appropriate use (TEA, 2009). Earlier research by Sherman (2008) found that an overwhelming number of Texas school superintendents supported the use of benchmarking in their districts, but Sherman's research did not isolate the use of released TAKS tests as the form of the benchmark. The purpose of this study went a step further; that is, before we can fully

understand the extent of what Nichols and Berliner (2007) call "collateral damage" (p. xvi), we needed to know if these two practices were the norm in Texas classrooms or the exception. Anecdotal evidence supplied by listening to teachers—both at physical and cyber gatherings—suggested that such use was widespread, but a more scientific approach and some specific data was warranted.

1.6 Approach to Research

I decided to use the area group of English/language arts supervisors (The professional organization) as the population sample for the administration of the survey. As a group they represented districts across the north central Texas area. Those districts are both, large and small; urban, suburban, and rural; wealthy and poor. I believed that the respondents (a) had an understanding of TAKS and the data it produced; (b) had knowledge of the practices of benchmarking, data-analysis, and decision making in their districts; (c) represented a valid sample of Texas school districts, and (d) would answer the survey questions honestly.

1.7 Significance of the Study

My study is significant in that it (a) demonstrates the prevalence of the practice of using a released TAKS test in a diagnostic fashion (as a benchmark test) in the school districts of the sample group even though the test was not designed for such a purpose; (b) demonstrates the extent to which TAKS scores are used in isolation to make important decisions in the school districts of the sample group, again contrary to test design and intent; and (c) provides areas for further research to examine the effects of

these practices. Additionally, the results of this study will be timely as Texas moves to STAAR as it will inform the discussion on what were the benefits/deficits of the TAKS era. Furthermore, this population can serve as a representation of Texas as it contains a variety of school compositions. Since educational trends in Texas have had enormous impact on the national education scene, this study can also be utilized in the national conversation on educational reform.

1.8 Basic Assumptions of the Study

I believe that the practices of using TAKS data inappropriately—both scores and released tests—are widespread throughout Texas. Sherman (2008) found that 78% of school superintendents supported the use of benchmark testing, and many of those testing scenarios utilized the released TAKS tests. However, Sherman's research dealt more with the concept of benchmarks rather than the composition of those assessment instruments. This study makes no statement on the validity or worth of the concept of benchmark testing; it focuses only on the use of released TAKS tests as instruments in benchmarking. Driven by the demands of state and federal mandates, administrators at both the campus and district levels are using this data, but because of what Popham (2008) called "abysmal assessment literacy" (p. 125), they are using it in ways that run counter to the test design and that can lead to what Nichols and Berliner (2007) call "collateral damage" (p. xvi).

1.9 Procedures for Collection of Data: Design and Rationale

After securing all necessary permissions and fulfilling all protocols, I moved to the active research phase of the study. The professional organization group meets monthly throughout the school year. I attended one of those meetings to explain my research study and requested their participation in a short survey on SurveyMonkey™. The questions (Appendix A) were designed to assess the nature of TAKS data (including released tests) used in their respective districts.

Data gathered does not identify individual districts by name, but rather by category (urban, suburban, rural, size, etc.). I believe this was an important factor in assuring the veracity of responses. No district wants to be singled out in a negative manner (and possible responses to the questions that can be perceived as either positive or negative), and no employee wants to be on record as the one who brought that attention to the district. Scholarly research is generally not thought of in the same light as, say, *The Dallas Morning News*, but recent events have made all educational researchers appear in a more suspect light. Participants in Arizona ELL research were guaranteed anonymity by the researchers, but later those research records were subpoenaed by the courts, resulting in a firestorm of concern in schools and the research community (Zehr, 2010). In California, *The Los Angeles Times* published the names of some 6,000 teachers and labeled them *effective* or *ineffective* in a so-called value added analysis (Felch, Song, & Smith, 2010). Their results were obtained by matching an economist's blind listing to a list of teacher names and campuses. Because of these and

other similar events, school districts and individual teachers are more cautious regarding their participation in research than previously, and guaranteeing anonymity was a necessary element.

I view this research as foundational, that is, it will be the catalyst for future research that looks to examine the specific effects of the utilization of TAKS data on a campus or in a school district. In ascertaining the prevalence of these practices, questions will arise concerning the "assessment literacy" of those who make decisions about students, teachers, and districts, leading—it is hoped—to further research.

1.10 Procedures for Collection of Data: The Population and Procedure

As previously mentioned, the population for this study was a professional organization, an affiliate of the statewide organization. This group of English/language arts supervisors is made up of district level administrators who oversee the ELA programs in their districts. At one of their monthly meetings during the 2011-2012 school year, I asked to be on the program to briefly discuss my research and recruit them to take part in a short online survey. Because the membership roster was rather small, I also added ELA specialists from outlying districts to build up the sample size.

1.11 Procedures for Analysis of Data

SurveyMonkey™ was chosen for the ease of use by the respondents and also for the manner in which the data is returned to the researcher. I took that data and used the appropriate statistical measure to analyze the data to determine if that data set provided

answers to my research questions. Issues related to the methodology of my research are discussed in Chapter 3.

1.12 Goals of the Review of Literature

In chapter 2, I present a review of literature that illustrates the complex nature of the problem of the data-driven concept applied to education. The problem is multi-faceted and resists quick analysis and solutions. This introductory chapter has laid out the problem as follows:

1. Reform efforts of educators over the past 30 years have resulted in demanding accountability from schools, and that accountability since NCLB has been in the form of state-level high-stakes tests.
2. Testing generates copious amounts of data.
3. Due to what Popham calls "abysmal assessment literacy" skills of educators, this data is frequently misunderstood, misused, and misapplied.
4. Because of these misuses, harm ("collateral damage" as Berliner terms it) occurs to students, teachers, schools, and the country.

In chapter 2, I examine how we arrived at this place, the reason for the déjà vu feeling, why data cannot be used in the manner that NCLB/reformers demand, and the degree of harm caused. It is a complex story, dating back to the latter part of the 19th century, so this chapter is lengthy--but essential--to providing sufficient background to address the research question in a robust context.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Theoretical Framework

Education is a complex construct that defies simple explanation. Educational reform efforts, including NCLB, have often sought broad and simple solutions to address issues yielding limited effect. The current educational landscape has been created from those reform efforts and their effects, both the successful and the less so. I believed it important to gain a historical perspective of these efforts and decided to use the theoretical framework of Anthony Gidden's Third Way theory of government and social change as expressed in Hargreaves's and Shirley's (2009) *The Fourth Way: The Inspiring Future for Educational Change* as my philosophical framework for a review of literature. This approach contextualizes the changes in education in the United States since World War II by describing them as either *First, Second,* or *Third Way*. This approach gives a clear and detailed description to the question "How did we get here?" Hargreaves and Shirley propose a way for the future of educational change—*The Fourth Way*—a combination of the best of each era. When viewed through this lens, the evolution of American education in the latter half of the 20th century is revealed to be often reactionary and rarely as innovative as is claimed. In truth, we have been here before but failed to learn from the mistakes and shortcomings of past endeavors.

Table 2.1 From the First Way to the Third Way of Educational Change

	The First Way	The Interregnum	The Second Way	The Third Way
Control	Professionalism	Professionalism and bureaucracy	Bureaucracy and markets	Bureaucracy, markets and professionalism
Purpose	Innovation and inspiration	Quest for coherence	Markets and standardization	Performance and partnership
Trust	Passive trust	Growing suspicion	Active mistrust	Public confidence
Community engagement	Mainly absent	Parent communication	Parent choice	Delivery of services to communities
Curriculum	Inconsistent innovation	Broad standards and outcomes	Detailed and prescribed standardized	Varying prescription with increased coaching and support
Teaching and learning	Eclectic and uneven	Prescriptively driven by standards and testing	Direct instruction to standards and test requirements	Autocratically data driven yet customized
Professionalism	Autonomous	Increasingly collaborative	Deprofessiona-lized	Reprofessional-ized
Professional learning communities	Discretionary	Some collaborative cultures	Contrived collegiality	Data driven and professionally effervescent
Assessment and accountability	Local and sampled	Portfolio and performance based	High-stakes testing and testing by census	Escalating targets, self-surveillance, and testing by census
Lateral relations	Voluntary	Consultative	Competitive	Networked

Note. From *The Fourth Way: The Inspiring Future for Educational Change: The Inspiring Future for Educational Change,* by Hargreaves and Shirley (2009, p. 44). Copyright 2009 by Corwin. Reprinted with permission.

In this chapter, I take an in-depth look at each of these stages of American education with an emphasis on the elements that can be directly seen as having impact

on the current educational climate. It is a somewhat lengthy effort, but one that is necessary to break out of the cycle of repetition that plagues educational reform. Hess (2010) captured the situation in the title of his book examining how reformers continue to do the same things in an endless loop: *The Same Thing Over and Over* (ironically, Hess does not include himself in that loop, even though he proposes many of the same ideas.) The following table, taken from Hargreaves and Shirley (2009), charts the course of educational change from the First Way to the Third Way and offers a partial progression/road map of this chapter.

While the aforementioned table clearly shows movement, in this chapter, I also show there is no unilateral movement through these phases. It can be safely asserted that the First Way days are long gone, but NCLB nationalized and entrenched Second Way policies so that the promise of the Third Way was cut short by a Second Way resurgence. Even the election of Barack Obama in 2008 which promised broad change, did not bring change in the field of education. Obama's educational policy embraces the Second Way rationale of the Bush administration policies and propels it even further. (Ravitch, 2011)

Sharing a platform with Andy Hargreaves at the Texas Association for Supervision and Curriculum Development (TXASCD) annual conference in Houston in 2010, Ken Kay—then head of the Partnership of 21st Century Skills—asked Hargreaves how we could be ready for the Fourth Way when America had not even reached the Third Way yet. Kay was making an important point: the Second Way thinking

manifested in NCLB is firmly fixed in American education, and the current crop of "reformers," such as Michelle Rhee, Joel Klein, Arne Duncan, and Bill Gates, push for an even greater Second Way manifestation.

In this chapter, an additional element is to privilege English/language arts (when appropriate) in considering the impact of these practices and policies. As a secondary reading specialist, my primary concern is the development of the literacy level of adolescent students. The implementation of Second Way ideas has caused, in Berliner's (2007) words, "collateral damage," and the research detailed in chapters 3 and 4 of this work focuses on determining the extent of several damaging practices in North Texas ELA classrooms.

2.2 The First Way: Idealism or Chaos?

The First Way is characterized by government structure with little or no accountability, leaving much room for innovation and creativity. It cultivated optimism that great problems (such as poverty, as one example) could be solved through government action, and that in doing so, society was bettered, and those bettered in turn benefitted the economy by "developing pools of talent that would fuel future prosperity" (Hargreaves & Shirley, 2009, p. 3). The First Way was also a tenet of the economic ideas of John Maynard Keynes, who greatly influenced the post-WWII governments of Europe and the United States, and whose ideas are experiencing a resurgence (as an antidote to recession) after falling out of favor in the 1970s (Fox, 2010).

Hargreaves and Shirley (2009) describe the First Way with the term "welfare state" (p. 3). Their use of "welfare" is broad, relating more to societal structures (including education) that support the well being of citizens. The First Way roughly covers the time period from the end of World War II to the mid 1970s. After the upheavals of the Second World War, "the social safety net of the welfare state appealed to war veterans of all ranks and their families. Having made so many sacrifices, they now wanted the opportunities and freedoms for which they had fought" (Hargreaves & Shirley, 2009, p. 3). The G.I. Bill of 1944 was a manifestation of this ideal: the government provided money and opportunity, and veterans exercised their hard-won freedom of choice to pursue higher education or vocational training, buy homes, start businesses, and participate in other forms of upward mobility. The impact of the G.I. Bill cannot be underemphasized; it completely transformed the landscape of America (U.S. Department of Veterans Affairs, 2011).

However, the word "welfare" carries an unfortunate connotation in the United States and immediately conjures up images of undeserved handouts. From the mid-1960s onward, stories of welfare fraud captivated and infuriated Americans. In 1976, then-presidential candidate Ronald Reagan made an issue of a "welfare queen" who illegally obtained well over a hundred thousand dollars—tax free—a year (Gilliam, 1999, p. 2) This story and others (the "Welfare Cadillac") have become part of American mythos and are still widely believed today, even after they have been disproved and after significant Welfare reform during the Clinton administration

(Gilliam, 1999). The word carries baggage with a negative connotation and is synonymous with "Big Government," another loaded term. Reagan was addressing reservations that were becoming more prevalent among conservatives. Reagan was seeking an alliance with business to undo the landmark legislation of the First Way: Lyndon Johnson's Great Society programs, which included the Head Start Program and the Elementary and Secondary Education Act (ESEA), according to Adler (2010). Many are not aware that NCLB was, in fact, the 2001 reauthorization of ESEA, which faces reauthorization every 5 years. The current reauthorization has been postponed since 2007, and because of the negativity attached to NCLB, will most likely resume its old name, ESEA. President Obama has challenged Congress to have the reauthorization complete before the start of the 2011-2012 school year; but as of April 2012, the legislation has not advanced (Klein, 2011). Reagan's timing was not incidental: by the mid-1970s, Second Way ideas were emerging and taking root as the public grew dissatisfied with perceived shortcomings of the First Way approach.

2.1.1 The First Way: In the Classroom

Teachers who were in the classroom in the 1960s and 1970s view the period through nostalgia-tinted eyes. Hargreaves and Shirley (2009) explain that the longing for the days when teaching was a respected profession with "high levels of trust" and educators felt "they were left alone to get on with the job," has two distinct variations (p. 5). Since teachers were so autonomous, those who taught in more "innovative schools" remember the freedoms they had to develop classes and materials to meet the

needs of their students. They viewed their mission as preparing students "to change the world" (Hargreaves & Shirley, 2009, p. 5).

The First Way approach to education was evident in my high school experience in the mid-1970s. The course offerings were more like a college class catalog: I could (and did) choose from classes like Contemporary Social Issues, Following the Frontier, Ancient and Medieval History, and American Political Development in social studies. In English, I could choose from Humor and Comedy in Literature, Shakespeare, Mythology, Fundamentals of Writing, Science Fiction, World Literature, American Novels, and Poetry. Science offered Advanced Physiology (a must for future doctors), Ecology I and II, and Horticulture I and II, as well as the traditional biology, physics, and chemistry. I began high school the week after Nixon resigned, and history classes were full of debate on the role of journalism in a democracy, abuse of power, and, of course, how we could make our system better (we were going to change the world, after all). The curriculum was what the teacher chose it to be. Students were taking those lessons personally. That first week, hundreds of students walked out of first period to stage a sit-in in the foyer to protest against a more aggressive tardy policy. I was thrilled to be attending "the most liberal high school in the state"—granted, a very conservative state that ultimately became even more so. I was thrilled beyond imagination at the opportunities I had and the relevance of my high school experience. I did not realize then that I was observing (in the mid-1970s) what were the death throes of the First Way approach to education.

On the other hand, Hargreaves and Shirley (2009) note, teachers who taught in more traditional schools in that era remember they had the freedom to teach the mandated courses (the traditional English I, II, III, etc.) in ways they preferred, including long lectures where students took copious notes. They recall "schools that were smaller, where unmotivated students left early for employment, and where the students who wanted to stay wanted to learn" (Hargreaves & Shirley, 2009, p. 5). Those students who could not handle school because of academic deficiencies or behavioral issues dropped out and went right into the work force. In short, the innovative school had the college course catalog and the more free-wheeling college discussion approach; the traditional school had the college teaching approach: but what they had in common was the freedom to choose an approach.

Teachers, like most people, have a longing for the "good old days." At the heart of this nostalgia in education is the concept of the teacher as professional in the sense that the professional sizes up the situation and calls the shots. The professional determines what to teach and how to teach it. The professional evaluates the success or lack thereof in his/her endeavors. Teachers today, even those who came to teach long after First Way thought had been shelved, keenly feel this *recherché du temps perdu*. In extreme cases, today's teachers are told what to teach and how to teach it to the tiniest detail. In 2010, the Sixth Circuit Court of Appeals upheld the nonrenewal of a teacher, Shelley Evans-Marshall, who was let go for her choice of texts in her high school English class. She maintained that as a professional, she had the academic freedom to

select the texts for her class using her judgment to determine the works that could best achieve the educational goals she was hired to teach. It was, she asserted, a matter of free speech. The court ruled that, in effect, her speech had been "hired" by the Board of Education and that she did not have the right to make those curricular choices (Walsh, 2010).

The decision by the Sixth Court of Appeals illustrates how far education has come from those days of First Way thinking regarding classroom practices. The teacher, in general, is no longer viewed as a wise professional who makes decisions based upon knowledge and experience: the teacher is an employee, doing what the teacher has been told to do. Substituting teacher judgment for district control can be considered insubordination. This is far removed from the attitude of the First Way. For example, consider the writing process movement of the 1970s, which created an emphasis on "process instead of product," as Murray wrote in 1972. That placed emphasis on the student rather than the finished product, and that idea as refined by Murray, Graves, Newkirk, and especially Atwell, redefined how writing was taught for a generation. It was the connection of expert teacher writer to novice student writer that was the spark for student learning. (In effect, the teacher stance was, "I am a learner, just like you. Let me show you how I write and I will help you develop your own style." The teacher's role is essential: the master, the mentor, the guide. Contrast this with other models where the teacher explains writing by the book or to the test. The teacher writer is not an essential part of that vehicle as we shall see.)

In true First Way fashion, accountability for this was determined not by any external educational protocol or structure; accountability was determined by the teacher—the professional—who oftentimes utilized "real world" accountability structures such as letters to city governments, poetry for Mother's Day presents, presentations to planning boards, etc. (Atwell, 1987). Accountability—and curriculum—were not set at state levels and then imposed on classrooms. The classroom was a world unto itself. Atwell encouraged teachers to go into their rooms and "shut the door" and create an environment where student reading and writing flourished and then "open the door" and invite others in. (1987)

When contrasting the days of the First Way with today's stressful environment of high-stakes testing and accountability, teachers "grieve for the passion and creativity that had been taken from their teaching" (Hargreaves & Shirley, 2009, p. 5). Passion and creativity were at the heart of many of the language arts professional development efforts of the era. The National Writing Project (NWP) began in 1974 with the California Bay Area Writing Project and spread nation-wide. While the goal of the project is to increase student achievement—the same goal as all Second Way efforts—the path to this goal was through professional development of teachers. NWP views "teachers as the best teacher of teachers" and, during the required summer institute, develops the teacher as writer. (After all, how can one teach what one does not do?) Passion for writing, and creativity in conveying this passion to students is paramount. The thought is that by creating teacher writers, we create student writers. Contrast this

with the New Jersey Writing Project in Texas (now Abydos), which is "student-focused" on the teaching of writing, notably in how to be successful on standardized testing. Their website touts methodology based on "true scientific research" from the Educational Testing Service that is not based "upon conjecture, spurious results, or falsified numbers but upon the hard and fast rules of science" (Abydos website). That sentence is steeped in Second Way thought. A focus on "student results" sounds very logical and most appropriate; in truth, however, it discounts and marginalizes the role of the teacher.

Ironically, Second Way accountability for writing still incorporates writing process thinking; that is, those assessments include the steps of the writing process, but have so formalized them that the steps are less about student writing and more about items on a checklist. The ideas of Murray, Graves et al. dealt with the process and the writer. Developing the writer was the true goal, and the steps in the process were a concurrent, organic process. Teachers were strongly encouraged to write alongside their students and share their own processes. All genres of writing were encouraged, with none especially privileged. Students were given much control over their topics and writing speeds. First Way thinking was comfortable with that sort of ambiguity; as will be shown, the Second Way was not.

It is important to note that in First Way schools success was defined by the teacher primarily, or perhaps, some campus or district criteria. Success was measured entirely by local entities. They were the ones who decided if students had learned.

During these days, many schools gave such tests as the normed-reference Iowa Tests of Basic Skills, but the results remained a local affair and were not used in a high-stakes manner. There was no system of rewards or penalties for performance on the test; rewards and penalties varied from campus to campus and were based on other metrics.

2.1.2 The First Way: Death Knell

The downfall of First Way thinking can be summed up in one word: inconsistency. Some schools were good, some great, and some brilliant. Likewise, some were mediocre, some bad, and some horrific. (To borrow a phrase from Charles Dickens' *Tale of Two Cities:* "It was the best of times, it was the worst of times . . . it was a time much like today.") The autonomy that teachers so loved offered no guarantee of consistent performance across time or geographic space. With minimal or no oversight, schools were free to do whatever they wanted. In some cases, this led to truly innovative educational systems like my alma mater, but other schools did not fare so well. Many continued as they had always been in spite of changes in society and their own populations. Other schools were constantly searching for the newest idea: "Fads were adopted uncritically, and many young radicals turned schools upside down during their brief tenures before heading for greener pastures" (Hargreaves & Shirley, 2009, p. 5). Combined with the social upheaval of the 1960s, the war in Vietnam, Watergate, the Energy Crisis, and the economic recession of the early-mid 1970s, the public lost faith in the welfare state and began expressing a loss of trust with the education system in America. The days of autonomy were numbered.

2.3 The Interregnum: Vitality Before Conformity

Hargreaves and Shirley (2009) claim that between the First and Second Ways, there was a "complex and contradictory interregnum" (p. 6). Then President Reagan "infused market principles into the welfare state" to cut back on the role of the federal government (p. 6), The notion that schools should be run as businesses began to be seen in public discussion, and "citizens" now became "consumers" or "customers." Students (and more frequently, student learning) were "products" (p. 9). The respect and trust that were automatic for educators during the First Way became replaced with "active mistrust" (Hargreaves & Shirley, 2009, p. 9). And yet, the First Way refused to go silently. The charter school movement began (fueled by teachers fed up with the bureaucracy of the system; for-profit schooling was still decades away), and the concept of assessing students through portfolios began to develop and, for a brief time, even flourished. But there was no room for individualization anymore; increasingly, there was a broad push for "coherence" after the "inconsistencies" of the First Way (Hargreaves & Shirley, 2009, p. 7).

President Reagan was never a fan of the Department of Education; indeed, one of his campaign promises was to abolish it. He was never able to accomplish that, but he did commission *A Nation at Risk* (1983), which put the nail in the coffin of First Way autonomy by claiming that American education is failing because of "a rising tide of mediocrity that threatens our very future as a Nation and a people" (*introduction*). The take away from this was that American schools were broken. This claim has been

since revealed to be largely unsubstantiated, yet the misconception persists. The conclusions of the report have since been disproved in both other government reports (for one, the Sandia Report commissioned by the Department of Energy; Stedman, 1994), and the work of many scholars and researchers, including Bracey, Berliner, Biddle, and others. Ansary (2007) writes the [NAR] report was "misquoted, misinterpreted, and often dead wrong" (p. 1)

However, much like Reagan's welfare queen, this idea has entered the popular mythos and refuses to die. It is seen on the big screen in the documentary *Waiting for Superman* (2010), which superficially tags teachers as "good" and "bad" and uses a cartoon figure of a "good" teacher opening the heads of students to pour in information, while the "bad" teacher sits behind a desk reading a newspaper. It must be noted that however widespread this idea of a broken system, there is a peculiar twist as reported annually by the Phi Delta Kappa/Gallup organization's report on educational attitudes. Consistently, parents give schools, as a whole, low grades, yet the schools their children attend are always rated above average, with B or higher (Berliner & Biddle, 2006; *Phi Delta Kappan,* 2010). The public perception seems to be American schools are "broken," but not *my* kid's school. Do parents evaluate schools on factors in addition to test scores? Or, is it that numbers are not as conclusive as they might be perceived to be?

The *Nation at Risk* report launched many reforms, including The National Board for Professional Teaching Standards (NBPTS), which established national certification

of teachers as a plan to both improve teaching and increase the professionalism of classroom teachers. NBPTS is an interesting hybrid of both First and Second Way thinking. It follows the First Way thought of increasing respect and trust for teachers as professionals, and yet it also merges Second Way accountability via standards and measurements (NBPTS, 1987). Another area of reform involved curriculum, which saw the birth of the standards movement. National standards were written by subject-area professional organizations. For example, the National Council of Teachers of English (NCTE) and the International Reading Association (IRA) collaborated on the English Language Arts Standards. Most states voluntarily began writing state standards; NCLB's mandate for standards would pull in the stragglers.

2.4 The Second Way: Standardized and Market-Driven

The hallmarks of the Second Way are "government centralization and standardization of educational goals" (Hargreaves & Shirley, 2009, p. 8). Ironically, President Reagan who saw little role for the federal government in education virtually guaranteed an even greater involvement because of *A Nation at Risk.* If education were broken, we could not depend on the people who supposedly did the breaking to fix it, could we? (Ansary, 2007). Hargreaves (2009) sums up this promise that was made to "consumers" as one that "promoted a sense of urgency, attended to all students, increased teachers' skill levels" and that would serve disenfranchised populations by "boost[ing] and equaliz[ing] achievement" (p. 9). Because of the disenchantment and distrust that "customers" felt toward the former First Way institution, whose decisions

had led to this state of brokenness, accountability was now tied to a commitment to gather comprehensive and precise data on student achievement (Hargreaves & Shirley, 2009) p. 9). If previously success had been a rather nebulous claim with an unclear trail of accountable persons, things would now be precise and devoid of any sort of value judgment. The proof of success was now in data, not in any qualitative measure; and that data was very narrow and very specific.

The National Reading Panel (NRP) is the actualization of that idea. The Panel, convened in 1997, was charged by Congress to assess the effectiveness of various approaches in the teaching of reading. The Panel only looked at what they considered to be "scientific research" and dismissed all qualitative studies, as well as studies which were "experimental or quasi-experimental" (Allington, 2002, p. 4). In doing so, they rejected out-of-hand the studies that supported evidence that validated whole language approaches, interjected the phrase "scientifically-based research" into educational jargon, and carried the reading wars into a new cycle of "my research is better than your research" (Allington, 2002, p. 4). Critics of the panel's composition, work, and findings were ignored. The Panel's report, "Teaching Children to Read" was issued in 2000 and was used to craft Reading First, an initiative that was part of NCLB (Allington, 2002).

Reading First was later found by the Inspector General to have serious issues with conflict of interest. Some consultants hired by the Department of Education to train states and teachers were authors of programs that made a good deal of money from these connections (Feller, 2006). In a nationwide study in 2008, it was found that

Reading First did not improve students' reading comprehension (Department of Education, 2008). Now it was the turn of the National Reading Panel/Reading First groups to cry foul because they believed research supporting their positions had been marginalized or contaminated (Stern, 2008). After all this time, there is no clear-cut scientific (or otherwise) evidence that clearly establishes one particular approach in teaching students to read.

In 2011, Wolk, former editor of *Education Week* who has followed every step in educational reform over the past 30 years, finally put in print what many had long thought: there is research to back every side in educational debates. Wolk added that anyone claiming to have the definitive research is trying to sell something. He may be presenting a cynical viewpoint, perhaps, but it injects a giant dose of much-needed pragmatism into the debate. When differing parties in the education debate claim that research supports diverse and even contradictory ideas, what conclusions can legitimately be drawn? It is not surprising to see that research does not lead to one clear answer or approach; as has been said before, education is a very complex issue and is profoundly situational. However, this continual tit-for-tat approach from the federal government and "experts" is not helping to improve education. Research, like data, can be manipulated. An important lesson was learned from the Reading First debacle: follow the money. Who is profiting from the changes? Even beyond those considerations: what works in one locale may not work in a different locale (Wolk, 2011). Researchers and government agencies such as NRP seek one solution that can be

applied in all situations and guarantee positive results, but there are too many variables, too many factors that cannot be pre-considered. In short, teaching is not just a science—it is an art as well (Marzano, 2007). Wolk writes that a strong dose of common sense must be included when considering research to impact educational practice. (2011)

NCLB is the legislative expression of the promise that oversight and standardization would improve education. It was bipartisan, introduced with great fanfare, and touted as the way those broken schools would be mended. But those who live by the market shall die by the market, and the forces of standardization and uniformity that were to provide coherence instead produced both a straight jacket and even more inconsistency, only this time with copious and often questionable data. (Hargreaves and Shirley, 2009) NCLB attempted to define a complex issue (educational gaps) with a simple solution (accountability through high-stakes testing), and capitalized on the public's limited attention span by reducing the entire school structure and population (students, parents, teachers, administrators, community, etc.) into one number that purported to measure both learning and quality. Johnson (2007) found that while the public was concerned about education, they were not interested in the "nuances of the debate" (p. 36). Education is just too complex for that sort of thinking; but, Second Way approaches demand just that. Education is a business; business success boils down to profit, educational success boils down—they claim—to test scores: or rather, a test score. Multiple data sources bring additional considerations and

complications to the table that may confuse or dilute findings, so for accountability under NCLB one single test score is utilized.

This metaphor of school as a business does not fit education. It never has, but that does not dissuade policymakers—and those who stand to make a buck. Public education exists—and always had existed—not as a business, but as an instrument for the public good. (Ravitch, 2010)

2.4.1 The Second Way: Accountability

Like many reforms, the theory behind the accountability movement sounds logical. The job of schools and teachers is to educate students. As such, they should be accountable for student learning. In order to determine this learning, we need a way to measure the learning. That measurement needs to supply clear lines of accountability, but also show where there are gaps in both individual student learning and in the performance of groups of students in the system.

The term *data-driven* is omnipresent in education, but just what does it mean? There is no clear consensus, and one school's data-driven program can be vastly different from the school across town. Bambrick-Santoyo (2010), a leader in the data-driven movement, comments that it is meant to have schools focus on one question: "[A]re our students learning" (p. xxii). It seems simple enough, but in Bambrick-Santoyo's terms, it is often "misunderstood."

For some, a data-driven school is simply one that conforms to the dictates of No Child Left Behind legislation. For others, it is any school in which assessments

> are used. More ominously, some consider data-driven schools to be those that ride roughshod over genuine learning in a mindless quest to "teach to the test." (p. xxii)

Bambrick-Santoyo lists four "key principles" in being data-driven, including one on assessment that calls for schools "to create rigorous interim assessments that provide meaningful data" and then analyzing, taking action (teaching what the data indicates that students need to learn) and creating a school culture that actively supports these practices (p. xxvi). It is interesting, from a critical point of view, to note that the assessments in this case are created internally, rather than being externally mandated, and that Bambrick-Santoyo calls for plural interim assessments rather than a terminal (summative) external exam. This is how data-driven instruction is meant to play out, yet, in many cases, is not how data-driven is practiced.

Bambrick-Santoyo (2010) goes on further to list eight mistakes, which he calls "perilous pitfalls"—that can invalidate the use of data-driven ideas (p. xxvi). They are as follows:

- Inferior interim assessments (Unless these are quality measures, i.e., 'well-thought out and carefully written, effective analysis of student strengths and weakness is impossible.')
- Secretive interim assessments (A secretive assessment is to save money, so schools reuse the same measures and do not go over the tests with students or teachers)

- Infrequent assessments
- Curriculum-assessment disconnect
- Delayed results
- Separation of teaching and analysis
- Ineffective follow-up
- Not making time for data. (That is, compiling data but not taking the time to examine it and discuss the implications for teaching and learning.) (p. xxvi)

Bambrick-Santoyo (2010), also calls excessive analysis of year-end testing a "false driver." He mentions the oft-used metaphor of formative assessment as being physical, and year-end tests as being autopsies and, calls these analyses "a waste of time" (pp. xxxii-xxxiii). In Bambrick-Santoyo's thinking, these tests come too late to have any effect on student learning; after all, those end-of-the-year students have moved on. Copious amounts of data are generated in every conceivable category, but to what gain? The next time the teacher faces his or her classroom, it will be an entirely new set of students. Bambrick-Santoyo argues that schools need to use their efforts more wisely and focus on the interim assessments that can be directly linked to student learning.

The infatuation with those summative, high-stakes tests boil down to the argument made by Harris, Smith, and Harris (2011):

> The tests are too narrow, they don't measure achievement accurately, they are not objective (though they claim otherwise), they are distorted by a high-stakes

> emphasis, and they do not predict future achievement. In short, the testing system we have is not what we assume it to be. (p. 139)

These tests are not good enough to be used as the basis for interim assessments, they are not good enough targets for rigorous curriculum, they are not good enough for college and career readiness. The Fordam Foundation's (2007) *Proficiency Illusion* report confirms this, yet states and policymakers continue to insist on using these scores as the yardstick for measuring success and labeling/punishing failure.

Ideally, after an assessment, the results can be broken down to give information on which skills the student has mastered and which need attention. This is often called "laser focusing" or "targeted" instruction. The idea is that a teacher will recognize that a student does not understand a particular skill, can pull that student aside for remediation/ additional instruction, and ensure that he/she is "caught up" with the class. That is the theory, and as indicated earlier, it generates copious amounts of data with an almost limitless number of ways to express that data (charts and graphs based on demographics, teachers, question numbers incorrect, weak/strong student expectations, etc.). Any of these charts can be—and often are—presented with a sense of urgency. Once I observed a data-driven/obsessed administrator confront a middle school teacher over early benchmark data: "Aren't you at all concerned that 17% of your students did not answer this question correctly?" "No," the teacher deadpanned. The teacher was correct. Having 83% of students get the correct answer six months before the administration of TAKS is not something to be alarmed about, especially on a reading

test that measures the application of multiple skills in each question. To many, this scenario seems oddly familiar. We have, in fact, been here before.

2.4.2 The Second Way: Déjà vu All Over Again (The Mindset Behind Testing)

Before the First Way approach, before World War II, and even before World War I, there was the efficiency movement. The idea that schools should be treated as businesses and evaluated according to productivity is not new. Education has been here before. In his book *Holding On to Good Ideas in a Time of Bad Ones: Six Literacy Principles Worth Fighting For,* Newkirk (2009) accuses contemporary reformers of "presentism" and acting as if "the world were created sometime in the 1990's" (p. 12). The great flaw in this, Newkirk points out is that these reformers are unable to see "what is happening now as part of a pattern" (p. 12). In short, they are unable to learn from history and so, the adage goes, they are "doomed to repeat it."

Newkirk (2009) posits that the guiding ideas of NCLB have been in place for almost a century in this country and longer in Europe. Zhao (2009), associate dean for global education at the University of Oregon, writes that those same ideas have been at play in China for thousands of years. Ironically, China is moving toward a more Western approach at exactly the time the U.S. is moving toward their abandoned models. Zhao explains he is confounded by the American attitude toward Chinese education. .

The Department of Education was seemingly unaware of the irony when they selected and built a little red schoolhouse in front of their offices to symbolize NCLB.

Then Secretary of Education Rod Paige said its choice as the physical representation of the legislation was intentional: "We serve the ideal of the little red schoolhouse. . . . It is one of the greatest symbols of America—a symbol that every child must be taught and every child must learn" (as cited in Glod, 2010, p. 1). The irony is that the little red schoolhouse had little or no oversight, no reliance on scientific research or external structures, and that one autonomous teacher planned and executed the education of his/her charges based solely on his/her own experience and content knowledge. Monitoring was sparse with rare visits from a district superintendent or other official. Newkirk (2009) writes that his mother was a teacher in such a school during the Depression, and the major concern of her lone supervisor was her ability to tend to the fire so that it would not die out. However, 20th century America was moving away from agrarian communities to small towns and cities, and larger—more systemic—schools were replacing the one-room school.

Those larger schools were viewed as ripe for standardization (as seen by The Committee of Ten's recommendations for high schools in 1892; Newkirk, 2009) and the application of efficiency models. In the mid-19th century, Pennsylvania Superintendent of Schools William Harvey Woods laid out the case for the graded school, arguing that this plan demonstrated the same "division of labor that prevails in well-regulated business establishments, whether mechanical, commercial, or otherwise" (as cited in Newkirk, 2009, p. 15). Woods goes on to say that, while the "individuality" of the teacher should be respected, he makes it clear that the teacher exists to serve the

system, and must follow that system, or else he is "an unworthy member of the profession" (as cited in Newkirk, 2009, p. 16).

The division of labor attitude is still found in the so-called reform movement of today. Bill Gates, a de facto leader, took a very Woodsian stance when he answered Diane Ravitch's criticisms of his ideas with the *ad hominem* counter charge that she preferred "the status quo" of the presumed broken schools which obviously makes her (note: not her arguments) one of the "unworthy" members of the profession (Lyons, 2010, p. 52). Very little has changed. Newkirk (2009) frames Woods' intent thusly:

> In effect, Woods is claiming to possess *coercive knowledge*, a set of principles so universal, that any attempt by teachers to resist or reject them amounted to professional irresponsibility—a dereliction of "duty." These principles, he claims, "belong to every good system of instruction." And how did Woods divine these principles? In his preface he claims, in an odd passive sentence, that they have "been suggested by the author's diary of visits to the schools of Chicago and other cities." (pp. 16-17)

Wells (as cited in Newkirk, 2009) used his own opinions to set forth the standard of education that all teachers must follow or else be considered unfit. Substitute data for opinion and we see the early 21st century as *déjà vu* all over again, to borrow from Yogi Berra. Teacher and school effectiveness are determined by data, and as pointed out in chapter 1, the nebulous nature of data and the propensity for misuse (more on this shortly) make its absolute use less reasonable than the idea suggests.

Newkirk (2009) goes on to say that while Woods lacked the "authoritative tool" to establish the criteria for effective teaching, "science would be the answer" (p. 17).

The area of scientific management was birthed at Bethlehem Steel Yard in 1899 and revolved around the central tenet of "separation of planning and execution of labor" (Newkirk, 2009, pp.12-14). Frederick Winslow Taylor, the creator of the system, believed that the man who carried out the labor/task was unable to "understand the science of that trade" (Newkirk, 2009, p. 12). That knowledge was owned by a different man or men, who had a different education. Taylor (quoted in Newkirk, 2009) takes pains to point out that the second man's education is "not necessarily higher . . . [but] different" (p. 14). The first man's job was to do what he was told, the second man's job was to do the telling. Thus, was born the "cult of efficiency," and though its initial application was in the industrial setting, it soon spread to other aspects of life and eventually to education (Newkirk, 2009, p. 14). It is easy enough to see the logic in this method with, for example, building a skyscraper or dam. It is harder to see the logic where the "product" is not as concrete.

Even so, schools were a ripe field for this approach. In his book, *Education and the Cult of Efficiency*, Callahan (1962) quotes Elwood P. Cubberly, dean of Stanford's School of Education and leading educational thinker of the time, who wrote that,

> Our schools are in a sense, factories in which raw products (children) are to be shaped and fashioned into products to meet various demands of life. The specifications for manufacturing come from the demands of the twentieth-

> century civilization, and it is the business of schools to build students to the specifications laid down. This demands good tools, specialized machinery, continuous measurement of production to see if it is according to specifications, the elimination of waste, and a large variety in output. (p. 152)

Notice the dehumanization of the description. Children are raw material to be transformed into something else by mechanized teachers and schools systems that are custom building them to fit the demands of the consumer: society. Callahan cinches his argument with this quote from John Franklin Bobbitt, a leader in the efficiency in education movement who wrote, "education is a shaping process as much as the manufacture of steel rails" (p. 152). Looking back at First Way thought, it is clear that much of the extremes of the First Way were reactions to this cult of efficiency.

The writings of philosophical leaders Cubberly and Bobbitt and, indeed, the history of the efficiency movement in education, bring home a startling truth that is just as applicable to today's accountability movement. Decision-making, removed from the classroom and actual contact with students and put into the hands of managers—and especially those who have never been teachers—who look at spreadsheets and numbers to determine effectiveness, is corrupted and inhumane. To the numbers person, it makes no difference if the product is steel rails, widgets, boxes of cereal, or human lives. The science of mathematics is absolute. The idea that science is pure and trumps human experience ignores a fundamental truth of education: education is profoundly situational. What happens in one scientific study is not always (or even often) replicable

in other situations across the country, across age differences, across demographics, and perhaps not even across the hall or in the next period. The number of variables at play in any given classroom is staggering.

"Data doesn't lie," "You can't argue with the numbers," "We can only measure the quantifiable," and other sophisms are half-truths used to end debate in the accountability movement. Like the *ad hominem* attacks mentioned earlier in this chapter, they are meant to convey that opposition is futile. Theirs is the belief that there is one way, the one perfect way to accomplish all educational goals and that way is revealed through science; in this case, the science of mathematical measurement. Bobbitt, as cited in Callahan (1962), spelled out the idea that,

> When a method which is clearly superior to all other methods has been discovered, it alone can be employed. To neglect this function and excuse one's negligence by proclaiming the value of the freedom of the teacher was perhaps justifiable under our earlier empiricism, when supervisors were merely promoted teachers and on the scientific side knew little more about standards and methods than the rank and file. (pp. 90-91)

In other words, Bobbitt states that science will provide the one true way and, worse, that teachers and administrators who move through the teaching ranks to become administrators are too stupid to understand their work. It is no different from the attacks on teachers that dominate media coverage today. The concept of "data-driven decision making" is the modern equivalent of Bobbitt's "clearly superior message."

Examine the attacks on teachers and the same ideas are present. Mayor Bloomberg of New York recently appointed a publishing executive with no educational experience to head the school system. She lasted three months (MSNBC, 2011). The current Secretary of Education, Arne Duncan, moved from professional basketball player (in Australia) to running the non-profit Ariel Education Initiative, which was a division of the for-profit Ariel Investments (with $5 billion in assets on September 30, 2010 according to their website) to Deputy Chief of Staff of the Chicago Public Schools in 1999 to CEO of Chicago Public Schools in 2001. Along the way, he sat on the boards of philanthropic organizations, some with educational ties. The previous Secretary of Education had a background in journalism. The George W. Bush Institute at Southern Methodist University is developing a "fast track" to train private sector "promising leaders" (like Black and Duncan) as principals (Stahl, 2010). What those principals who come from outside education, Arne Duncan, Cathie Black, and the de facto leaders such as Bill Gates lack is rank and file experience: they have never been responsible for directly educating anybody. They have never been responsible for a classroom, writing a test, holding a parent conference, dealing with disinterested or disruptive students, nor have they dealt with any of the myriad experiences that come with being a teacher, and that creates a huge credibility gap within the education community. However, when one subscribes to the efficiency model and accepts that the division of labor idea, it matters not one bit. Indeed, it appears that teaching experience is a liability for educational

leaders in this model. Bobbitt's argument can be interpolated to say that these people from outside education know the science and are not contaminated by experience.

Yet the totality of Bobbitt's career and writing reveals puzzling contradictions and serious questions about the applicability of his ideas to today's situations. Bobbitt was, as indicated, an advocate for efficiency in schools, but he also was an early pioneer in the development of curriculum as we know it. He argued that classical curriculum should be replaced with a curriculum that fits the needs of the individual to fit into the industrial society. Bobbit's ideas were shaped by his experiences in the Philippines where he was on a team to develop a local curriculum. The team began by adapting American textbooks and curricula to fit the situation before realizing that it was not successful. What they needed to do—and ultimately did—was to develop a product that fit the needs of the people, at this place, at this time. Bobbitt wrote his team needed to educate individuals according to their capabilities. In Bobbit's words,

> This requires that the material of the curriculum be sufficiently various to meet the needs of every class of individuals in the community, and that the course of training and study be sufficiently flexible so that the individual can be given just the things that he needs. (as cited in Callahan,1962, p. 269)

Those are goals that any First Way thinker would endorse, but Bobbitt viewed them only through the lens of what society/industry wanted from students—(the need) rather than vice versa as a First Way thinker would. The warning comes from the last phrase "just the things he needs." Who determines what is needed? Not the student—the

demands of society. Society does not demand that schools produce students to change the status quo.

For example, Bobbitt, writing in 1912, did not believe that girls should be educated alongside boys because their needs were so "different." The curriculum for girls—in the Bobbitt system—involved training and education for girls only to assume the traditional roles of women in 1912 (as cited in Callahan, 1962). We can extrapolate this argument to see what would have been a very limiting education for African-American children, children with disabilities, children from low SES backgrounds, immigrant children—in short, we can authorize a substandard education for anyone who is not a white male using Bobbitt's framework. The system will always seek to maintain the status quo, but our Constitution challenges us to seek a "more perfect union" so reform measures then and now deserve a more critical look. It is not enough to serve a system; the system exists to serve the people.

Another tenet of the efficiency movement that is highly visible in today's educational climate is the deprofessionalization—and at times, downright vilification of teachers. Remember that in the factory model, the teacher assumes the role of the laborer who does not understand the totality of the project. Bobbitt writes that teachers "cannot be allowed to follow caprice" (as cited in Callahan,1962, p. 90), using the word "caprice" where professional judgment should rest. The experience and education of teachers is devalued in this model, and is characterized in pejorative terms. Much of the accountability movement seeks to reign in teachers who were doing whatever they

wanted, acting as lone agents rather than following a set curriculum or scope and sequence. This viewpoint suggests that teachers act irresponsibly in defiance of obvious truth. Newkirk (2009) writes that, "[t]eacher knowledge is portrayed as almost childish willfulness" when it is "contrasted with the solid knowledge of science" (p. 18). Newkirk goes on to explain that such a belief in science is called positivism, and it assumes there is a "uniformity of nature (including human nature)" and that science is capable of finding the ways to work through all the variables and locate "the bedrock of universal truth, to the laws of learning" (p. 20). However, truth—like data—is not an absolute term.

2.4.3 The Second Way: International Perspectives

Competition is an inherent part of Second Way philosophy, and nowhere is it more evident than in international comparisons. Because these comparisons fuel much of the reform movement, this section of the review deals with what is behind those comparisons. Since the publication of *A Nation at Risk,* Americans have been convinced that students in other countries "outperform" our students, and therefore, our educational system is inferior. Subsequent releases of Trends in International Math and Science Study (TIMSS) and Profile of International Student Achievement (PISA) data continue to co contribute to this idea. China, in particular, is held up as our main rival. In December 2010, when a snow emergency in Philadelphia postponed an NFL game between the Minnesota Vikings and the Philadelphia Eagles, the governor of Pennsylvania complained that,

> [W]e're becoming a nation of wusses. . . . The Chinese are kicking our butt in everything. If this was China, do you think the Chinese would have called off the game? People would have been marching down to the stadium, they would have walked and they would have been doing calculus on the way down. (Florio, 2010)

In his book *Catching Up or Leading the Way: American Education in the Age of Globalization*, Zhao (2009) explains that he set out to write a book about education reform in his native China, but,

> [R]ealized that what China wants [in education reform] is what America is eager to throw away—an education that respects individual talents, supports divergent thinking, tolerates deviation, and encourages creativity; a system in which government does not dictate what students learn or how teachers teach; and culture that does not rank or judge the success of a school, a teacher, or a child based on only test scores in a few subjects determined by the government. (p. vi)

Zhao is saying something truly remarkable here: America's main competitor wants to be more like us, and we want to be more like the model China is abandoning.

Zhao attended schools in China and taught English before immigrating to the United States in the 1990s. He is currently an educational researcher and holds the Presidential Chair and Associate Dean for Global Education at the College of Education at the University of Oregon. Additionally, his children attended American schools, and it was the "broad damage inflicted by NCLB as well as the growing enthusiasm for

more standardization and centralization" that caused him to change his writing direction (p. vii). Zhao (2009) explains the irony of the situation is that, while America is struggling to be more like China, China is abandoning that model because it realizes the limitations and liabilities of a high-stakes testing culture. Chinese educational reform consists of "an unwavering desire to undo the damages of testing and standardization" (p. vi). American educational reform calls for more standardization and more testing. Zhao admits he is perplexed by the situation. Furthermore, he contends that such trends in American education would be "disastrous" (p. x).

One of Zhao's (2009) chapters is titled "Why China Hasn't Beaten Us Yet." In it, he quotes Premier Wen Jiabao who told of a conversation with Qian Xuesen, a rocket scientist who had studied at MIT and earned his Ph.D. from the California Institute of Technology before returning to China to work on China's early space programs. According to Quian,

> One of the important reasons that China has not fully developed is that not one university has been able to follow a model that can produce creative and innovative talents; none has its own unique innovations, and thus has not produced distinguished individuals. (as cited in Zhao, 2009, p. 65)

Wen goes on to add that this is a great concern because while China's university population has been increasing, there has not been evidence of what Quian called "distinguished individuals" or true masters of their studies (Zhao, 2009, p. 65).

Zhao, 2009) calls this a "shortage of creative and innovative talent" (p. 65). He directly attributes it to the Chinese educational system, which has always depended upon high-stakes testing and standardization: the two trends in American educational reform that most alarm him. Indeed, those are the issues that also concern the Chinese educational reformers because they are seeking to undo the damage of those ideas (p. vi).

China is the "world's factory" where things are assembled cheaply, but they are products that are designed or invented elsewhere. The massive growth of China's economy is largely due to the "vast and cheap labor force" rather than knowledge or innovation, and cheap labor alone does not reap the full financial benefit of commerce. (Zhao, 2009, p. 65). The subtitle of Zhao's chapter is "The Costs of High Scores," where he explains how it plays out in China's economy.

> Despite China's astounding double-digit growth for more than two decades, its economy remains one that is labor intensive rather than knowledge intensive. The growth has been largely fueled by its vast and cheap labor instead of technology. In other words, as the "world's factory," China has been mostly making things designed elsewhere (Shenkar, 2006). According to a report of the Chinese National Statistics Bureau, only about 2,000 Chinese companies owned the patent for the core technology used in the products they produced in 2005; that number represents less than 0.003 percent of all Chinese companies in that year. (Zhao & Wu, 2005).

As a result, although products worth billions of dollars are made *in* China, they are not made *by* China. And that costs China. Zhao gives the example of the Bratz dolls that will earn a Chinese worker 17 cents for the labor involved, but will retail for around $20. The big profits go to the company that owns the patent, MGA Entertainment, a company based in California. Even though China has seen an increase in patent applications, the overwhelming number of factories still produce the innovations of others. While China's economic growth over the past few decades has been astounding, the Chinese leaders are "aware of the danger of an economy dependent on cheap labor instead of technology" (Zhao, 2009, p. 67)

Zhao (2009) is not the only one calling our alternately rose-colored and alarmist views of China into question. Oded Shenkar, who holds the Ford Motor Company Chair in Global Business Management at the Fisher College of Business at Ohio State University, writes that China is poised to become a superpower in the 21st century, but must actualize what he calls "the dream of indigenous innovation," which has thus far proved elusive (as cited in Zhao, 2009, p. 6). He has tracked patent applications in both number and type, and finds that even with a concerted effort, China still has not "establish[ed], so far, an effective indigenous network of technological innovation" (p. 69). The leadership in China is calling for a change in the system, including extensive educational reform. Zhao (2009) reports a speech by Communist Party leader (and future president), Hu Jintao, in which he [Hu] calls for China to become a nation of "innovation" and cites concern that China is "significantly behind advanced nations" in

science and technological development (p. 68). It is extremely ironic that that exact phrase is used in the United States to compare educational test results.

In early 2011, *Businessweek* published an article by Vivek Wadhwa, who has researched Chinese education in studies at Duke and USC, with the title "U.S. Schools Are Still Ahead—Way Ahead" that decries the national alarm that greets announcement of any type of international comparison testing story. He asserts that while there are plenty of areas for improvement in American education, there is no need for the national "inferiority complex." The American system of education produces, for example, engineers that are far superior to those produced by Chinese or Indian schools. Those engineers, he claims, need at minimum two to three years after school to be able to perform the tasks the job requires and most are of such "poor quality" that they are "not fit to work as engineers (Wadhwa,2011, p. 1). They do not match the productivity of American engineers until after several years of additional training. He blames this on the schools systems that stifle innovation. What many perceive as a weakness in American schools (not requiring more homework, allowing too much time for games and social activities, being taught to "challenge norms" and take risks, too much emphasis on social skills, etc.), Wadhwa feels is a critical component in producing innovative thinkers. Both China and India are attempting to make changes in their systems to encourage that kind of thinking.

Nevertheless, to use Zhao's (2009) word, Americans continue to "glorify" the Chinese Educational System (p. 68). China is held up as system worth emulating. Bill

Gates said in the December 20, 2010 issue of *Newsweek* that, "[t]he Chinese, who have a 10th of our wealth, are running a great education system" (p. 52.) A steady stream of literature over the past several decades has touted Chinese educational success while downplaying negative factors.

In truth, Chinese education is much like American education: there are impressive aspects and troubling aspects that co-exist. It is an oversimplification to say that their superior test scores mean China has a better system than the United States. As we have seen, the successful test scores do not translate to success in the after school life or else China would be content and not seeking to cultivate innovation. Zhao (2009) points out that his critiques of the Chinese system should in no way be construed as an endorsement of the current American education system. Zhao is frequently critical of the system, and especially of NCLB reforms. There is much in American education that needs to be improved upon, but adopting a Chinese approach to education without thorough consideration is foolhardy. Beyond the crisis of "indigenous innovation," (Shenkar, cited by Zhao, 2009), there are two byproducts, in my thinking, of the Chinese educational system that demand serious scrutiny.

The first area of scrutiny is fraud. China's ascendency to superpower status is jeopardized by the widespread fraud that permeates education, especially higher education and the innovation community as well as society at large. Jacobs (2010) writes in *The New York Times* that the latest revelation (a well-known "practitioner of traditional Chinese medicine . . . who claimed to come from a long line of doctors" was,

in truth, an unemployed textile worker) had once again drawn attention to "what many scholars and Chinese complain are dishonest practices that permeate society, including students who cheat on college entrance exams, scholars who promote fake or unoriginal research, and dairy companies that sell poisoned milk to infants." Jacobs points out that no country is immune to fraud, and that all have had experiences with it, but that Chinese "fakery" in education and scientific research will jeopardize China's ability to rise as a true economic superpower.

As indicated earlier, China is very concerned with developing its innovation capital. One measure of innovation is publication in professional journals. A 20-month experiment with the use of plagiarism software on professional journals (in medicine, physics, engineering, and computer science) published by Zhejiang University found that nearly a third of all submissions contained high levels of plagiarized material (Jacobs, 2010). A government study revealed that a third of some 6000 scientists admitted they had "engaged in plagiarism or the outright fabrication of research data" (p. 9) Another study conducted by the China Association for Science and Technology (2010) found that 55% of scientists in the study sample (some 32,000) said they knew someone involved in this type of fraudulent activity.

Of course, this attitude does not exist solely at this level of academia. Jacobs (2010) tells of Centenary College in New Jersey, which has satellite campuses in China and Taiwan. In July, 2010, the college shut down business schools in Shanghai, Beijing, and Taipei after discovering the extent of academic dishonesty among students. Some

400 students had their degrees withheld. When they were offered the chance to receive their degrees by taking a different exam, 398 students declined (Jacobs, 2010).

This culture of cheating intensifies in high school because of the pressures and high-stakes involved in testing and admissions to "good" schools. The pressures on students are "unrelenting" (more on this shortly) and cheating is viewed not as a form of fraud, but as a time saving device, and one that is necessary to be successful according to Jacobs (2010). Campbell's Law (the more importance a measurement, the more likely it is to be corrupted; this term is fully explored later in this chapter) is in full effect here: the stakes on those exams are so high, that corruption has become automatic and commonplace. Jacobs calls the student response "nonchalant" and quotes Arthur Lu, now a graduate student at Stanford, who said "[p]erhaps it [is] a cultural difference, but there is nothing bad or embarrassing about it"(p. A1) Technology has assisted students with both gadgets that covertly communicate such as wristwatches or pens with cameras and access to writers who will write papers for the students (Jacobs, 2010). All of this casts doubt on China's efforts to truly reach its potential as a world power and should serve as an important consideration in American efforts to emulate China's educational system.

The second area of concern is the mental health of Chinese students. Cultural expectations are high and the pressure is intense. Everything rides on success on the exam to get into a good school. Students attend school, and after hours and on Saturday, attend a "cram" school. Chen Weihua editorialized that "[t]he making of superb test

takers comes at a high cost, often killing much of, if not all, the joy of childhood" (as cited in Stack, 2011, Chap. 1). Stack also commented that because of China's famous "one child" policy, the pressure—that of generations—is enormous and takes its toll. One school that did especially well in the recent PISA scores celebrated that fact, but the vice-principal noted that the students "suffer very poor health, they are not strong and they get injured easily" and called for a concerted effort from all parties to scale back on the pressure students face (Stack, 2011, Chap. 1). The Chinese Association for Mental Health reports suicide as the number one cause of death for the 15-34 age group (Reynolds, 2010).

Kwan (2010) researched life-satisfaction among students in Hong Kong, which has some complex dynamics in that the former British colony is now part of a communist system, yet always had the Chinese cultural heritage and emphasis on educational outcomes (i.e., heavily test-oriented, getting into "good schools," etc.). Kwan found that the life-satisfaction was far lower in Hong Kong than in the United States, especially in the area of school-satisfaction. Family relationships, which also rated low, were tied to educational satisfaction. Kwan's research differentiated by gender, and he connected higher rates of male dissatisfaction with the increased cultural pressure placed upon males by parents.

Once again, it is important to note how complex education is, and how that complexity deepens at exponential rates when we attempt to make international comparisons. There are so many factors in play that any sort of comparison is difficult

to sustain after moving past the numbers. Even highly respected sources such as Merrow (2010) throw the numbers about with abandon, all claiming this "gap" in scores is cause for alarm. The value of the comparison is questionable.

Stories about American schools that are besting international comparisons do not make it to the evening news or pundit speeches. For example, *Education Week* writer Robelen (2010) reported that the Clayton School District, an affluent school district outside of St. Louis, took the PISA exam as a part of a study done by ACT. Their 15-year-olds, outscored every official PISA nation in reading and science, and placed second in math. Impressive results indeed, and ones that give validity to the claim that when test results are compared to American schools that have less than ten percent low SES numbers, American schools do as well as anyone in the world. However, what happens next in the article is very telling. Robelen goes on to damn with faint praise by saying such numbers would "be expected" from such a high-achieving campus and that it would be worth noting only if they failed to score high. He cites researchers from Stanford and MIT who also undercut the achievement by minimizing Clayton's accomplishment. The impression that American schools are inferior is so entrenched that even when presented with evidence to the contrary, belief in our own inferiority wins out.

2.4.4 The Second Way: Unintended Consequences

Second Way reforms produced a number of unintended consequences that are a part of the legacy of standardized testing and governmental oversight. As much as the

unintended consequences of the First Way led to its demise and transition into the next way, these unintended consequences are pressuring Second Way proponents to evolve. Is it possible that instead of ensuring that students are learning more, the opposite is happening? That standardized curricula (in some cases scripted even to restroom breaks and hand gestures) instead of being a guarantee of quality is actually a recipe for mediocrity? That accountability, instead of improving educational quality, has compromised the entire process, and many think it has done so.

2.4.4.1 Illusory Gains

Standards and testing were required of all U.S. states under NCLB, and those who were not already doing so scrambled to write standards and tests to be in compliance. In theory, it is a workable plan, but it soon became clear that there were major problems with this Second Way approach. First, since each state establishes its own standards, tests and proficiency levels, comparisons cannot be made between states. Is a student who is proficient in the State of Texas also proficient in the State of Massachusetts? Can a student be proficient in one state move to another state and then be found to be at-risk, unacceptable, or some other term for the bottom-of-the-barrel achievers? In its study, *The Proficiency Illusion*, the Fordham Institute (2007) found exactly that. By its own definition, the Thomas B. Fordham Institute is a think tank devoted to educational issues. It is generally thought of as a conservative organization, so one must take that into account when evaluating their message (Bracey, 2006). Fordham states, according to Bracey, unabashedly "[w]e believe in results-based, test-

measured, standards-aligned accountability systems" (p. 3). Therefore, the message of *The Proficiency Illusion* is all the more startling. The results of testing do not adequately inform parents of the reality of their child's educational progress. "Mr. and Mrs. Smith know that little Susie is 'proficient.' What they don't know is that 'proficient' doesn't mean much. This is the proficiency illusion" (Fordham, 2007, p. 2). By any definition, the word "proficient" implies success. Maybe just enough success, but success all the same, so there appears to be a large-scale fraudulent system of reporting student achievement.

The key is in the assessment, and how the results are interpreted. The test results need to be checked against a standard that is fixed. With fifty different state assessments based on different standards and with different scoring systems and scales, this is a critical point. One of the characteristics of a solid assessment is that a different assessment will produce the same results, confirming the validity of the first assessment (Seife, 2011). The report argues that in order to truly assess the assessments, a "solid and reliable . . . yardstick" is needed, and they chose the Measures of Academic Progress (MAP), an assessment developed by the Northwest Evaluation Association as that yardstick (Fordham, 2007, p. 3). The researchers then compared the state assessments and data thereof, with the 26 states that also had corresponding data from the MAP. They asked the three following questions:

- How hard is it to pass each state's tests?
- Has it [the test] been getting easier or harder since enactment of NCLB?

- Are a state's cut scores consistent from grade to grade? (p. 3)

The results shocked even the conservative Fordham group. Passing scores varied widely among states, ranging from the 6th percentile to the 77th. Additionally, the passing rates are not set in intervals, that is, the difficulty between grade levels is not consistent. It is far easier to be labeled "proficient" in elementary than middle school, giving parents the impression that their child is on track for success, when that may not be true at all (Fordham, 2007, p. 3).

The Fordham researchers call the entire testing system "unbelievably slipshod" and call this "big trouble" for anyone with a stake in U.S. education. The researchers state the obvious:

> America is awash in achievement 'data,' yet the truth about our educational performance is far from transparent and trustworthy. It may be smoke and mirrors. Gains (and slippages) may be illusory. Comparisons may be misleading. Apparent problems may be nonexistent, or at least, misstated. The testing infrastructure on which so much confidence has been vested, is unreliable at best. (Fordham, 2007, p. 3)

That unreliability is inherent in NCLB; they argue, because it left it to the states to individually set their standards, testing, and—this is important—own cut scores or definition of proficiency. Again, it must be pointed out that this observation comes from a fervently pro-testing, conservative group. It stands to reason that the anti-testing factions would consider the testing infrastructure faulty, but to read this condemnation

from such staunch testing-advocates gives cause for concern. While many students appear proficient, Fordham notes that, in Wisconsin for example, the cut score corresponds with the 14th percentile according to the MAP comparisons and that South Carolina's cut score puts those passing students at the 71st percentile (Fordham, 2007, p. 4). Could it be more inequitable?

It is. The report cites that many states are "internally inconsistent" with reading tests being far easier than math, and third grade expectations proportionally lower than eighth grade. Millions of parents, they claim, are being duped into believing their children are on track for success from elementary school when they are really heading for disaster in the upper grades. The contrary is true as well: middle school achievement may actually be much higher than reported. Likewise, math achievement may be higher and reading achievement lower (Fordham, 2007). The researchers could not determine with certainty because the system is not set up for such precision. Proficiency—and the lack of it—are illusory. The American Educational Research Association (AERA) warned in 2000 that, "[p]olicy makers and the public may be misled by spurious test score increases unrelated to any fundamental educational improvement."

While Texas earns kudos from the Fordham researchers for making their cut scores more challenging (the TAKS phase-in period gradually increased the percentage correct needed to meet standard), the cut scores examined for reading ranged between the 12th and 32nd percentiles (Fordham, 2007, p. 188). That is a very low threshold indeed. Additionally, these tests are not calibrated, with the 7th grade reading test

coming in as the most difficult. (This was in 2007; Texas made efforts to better align the tests in 2010.) In mathematics, the scores ranged between the 24th and 41st percentiles. Both of these are in the bottom half of the state comparisons, sometimes actually scraping the bottom. In the case of third-grade reading, only one state has a lower cut score than Texas (Fordham, 2007, p. 188).

It is scary to think that a third-grade student labeled proficient in reading is actually coming in at the 12th percentile—the bottom quartile—and would be actually targeted for assistance from programs like Title I. But it is the reality Texas educators face. Schools are under many pressures, but none more so than that of the state accountability rankings. When results are announced with great fanfare the first week of August, there is celebration at those schools whose proficiency levels are deemed high enough, but as the Fordham researchers point out, that proficiency is as much an illusion as the Emperor's clothes.

Politicians of all calibers, all levels, and both political parties love to claim "test scores are up," but determining the veracity behind that claim is difficult. Numbers may be higher, but do those numbers provide the proof that schools are better and that children are learning more than they were in the pre NCLB years? Can the numbers be trusted?

There are many who say no. Harvey (2004) writes that NCLB has become "the weapon of mass destruction" and that the accountability those test scores provide is only "illusory" (p. 18). Harvey chooses the same word the Fordham report did: *illusion.*

Egan (2005) writes that this testing is not even compatible with the mission of schools, and that the pursuit of this sort of measurement is counterproductive. Schmoker (2008) contends that it is possible for schools to show "steady gains" on such testing "without offering students intellectually challenging tasks;" indeed, Schmoker goes on to say that, "we discount evidence that fixating on data hindered instructional improvement in many schools" (p. 71).

In their book, *The Myths of Standardized Test: Why They Don't Tell You What You Think They Do*, Harris, Smith, and Harris (2011) systematically and with great attention to detail deconstruct the prevailing notions of the accuracy of standardized tests. They examine the creation of tests, how numbers do (and do not) work, and the types of decisions based on those numbers. The authors conclude that, "great harm is done by treating test results, which are gross and fuzzy indicators at best, as if their validity is absolute, their apparent precision is real, and their unexamined outcomes unchallengeable" (Harris et al., 2011, p. 63). Harris et al. also write that any test is subject to "random[ness] and unpredictable imprecision" (p. 189). Koretz (2008) also goes into detail explaining this lack of certitude about testing, as does Seife (2010) in explaining the margin of error inherently involved with standard testing.

A mini-scandal erupted in Texas in the summer of 2010 after TEA announced that test scores had hit their highest level yet. Reporters for *The Houston Chronicle* began investigating how TEA sets cut scores, that is, the point at which students can be considered to be successful on the test. Traditionally, failing school grades begin below

60% or sometimes 70%. It was a shock for some Texans to realize that passing rates on TAKS could be as low as 44%. (According to the Raw Score Conversion charts on the TEA website, www.tea.org) The article revealed that TEA determined the "passing" or "met standard" score only after all tests had been taken and scored. Only then did they determine the criteria for passing: after they had seen how students had performed, allowing them to in effect select the number of passing students they wanted. Adding to the speculation, 2010 was a gubernatorial election year.

Think of it this way: A teacher gives a ten-question test to the class. In a traditional model, only those students who had at least six questions correct would be considered passing. However, the teacher (for either base or altruistic reasons, it matters not) wants the maximum number of students to pass, so after looking at the results, the teacher sees that if he sets the cut score at 40%, only two students out of his class of 30 would fail. So he sets the cut score at 40% correct and can now boast of a 93% passing rate on his test. The newspaper report made this scenario clear, and readers—including a state representative and a school board president—were shocked to discover that cut scores were not the same from year-to-year, nor were they the same across subjects (Mellon, 2010). The article revealed to the general public (at least those who were paying attention) that passing the TAKS was not a purely mathematical formula, but rather a subjective decision where TEA could manipulate the numbers to have it appear that more (or fewer) students had passed. TEA maintains that is a quality control issue whereby they can adjust for difficulty of the test from year-to-year (TEA, 2010).

However, that explanation does not appear to be satisfactory when so much is riding on the outcome of the test. TEA should produce a consistent test that does not involve a moving target.

Another area of concern is cut scores vary across grade levels and subjects. The Raw Score Conversion Charts on the TEA website provide information on just how many questions students must score correct to meet standard. In some years, answering only 44% of the high school exit exam questions in social studies rated a passing score, yet to pass 5th grade science, a student needed to have 80% of the answers correct. The other subjects also vary widely.

Test scores are privileged over other data sources and accepted without question because of the presupposition that they are based on "objective" mathematics, but they should be examined as critically as any other product in the marketplace. Might we not ask for a second opinion? Seife (2010) points out that one of the qualities of a "good" test is that its results can be verified by a separate, independent test. What was intended to restore trust in schools has actually created a system of shortcuts, gaming, questionable moral calls, and—in extreme cases—illegal activity.

2.4.4.2 Gaming the System

Purported increases in testing as part of NCLB are under scrutiny as well. As mentioned, the *Proficiency Illusion* shows the inconsistent (and frequently low-balled) levels required to attain proficient rankings. Koretz (2008) asserts that any sudden gains are more the result of test prep strategies than any increased learning. Jerald (2003)

writes in an earlier *Educational Journal* article that schools are becoming skilled at gaming the system: that is, they have learned what the test demands and how to best the system. But that approach takes time and resources away from the non-testing aspects of education.

Elsewhere in this study, I discuss TAKS Writing and how the writing instructional program caters to the specifications of TAKS at the expense of the overall curriculum. The writing scores are some of the highest scores on TAKS. This should indicate a very strong writing program and college-ready students. The numbers of students requiring remedial writing classes would seem to suggest not. *The Chronicle of Higher Education* (2008) reported that nationwide, about half of incoming freshmen will need to take at least one remedial course, and "four of ten" will need a remedial writing class. In Texas, the numbers correspond (Hacker, 2011).

Each year also brings accusations of cheating by teachers and/or administrators who feel the pressure to deliver higher scores. These increases in state scores are also suspect because the advances are not mirrored in the National Assessment of Educational Progress tests (NAEP). Often called the Nation's Report Card, NAEP scores report only stagnant scores or very small gains; most of which were obtained from reforms put into place before NCLB. In the first half of 2011, major cheating scandals erupted in Atlanta, Washington D.C., and Philadelphia (Samuels, 2011). These may only represent the tip of the iceberg.

In 1975, Donald Campbell, a noted social psychologist, presented the concept that would later be known as Campbell's Law, that, "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 was intended to monitor" (as cited in Nichols & Berliner, 2007, pp. 26-27). When high-stakes are attached to a measurement, the more likely it is that the measurement will not only be corrupted, but also contaminate what is being measured. Nichols and Berliner claim that Campbell was able to predict the outcome of NCLB a quarter century before its invention: an untrustworthy/corrupted measure and a compromised learning system. This is what they consider the collateral damage of high-stakes testing.

2.4.4.3 Narrowed Curriculum

A frequent charge leveled against high-stakes testing/accountability is that the curriculum narrows to cover only what is on the test. In the larger picture, this means limiting or removing non-tested subjects, or within a subject, focusing on only what is tested, or teaching to the test. Popham (2001) explains this very clearly and conscisely when he writes that,

> [T]he first negative effect of today's high-stakes testing programs is that such programs divert educators attention from the genuinely important educational decisions they ought to be making. Thousands of American educators find themselves caught up in a score-boosting obsession that seriously detracts from their effectiveness in teaching children. The

> critical question of "How do we teach Tracy the things she needs to know?" is forced aside by this far less important one: "How do we improve Tracy's score on the high-stakes test she'll be taking?" (p. 16)

This narrowing is significant. Au (2007) analyzed 49 studies examining the relationship between curriculum and testing and found widespread trends of reducing curriculum to "testable pieces" (p. 262). As presented elsewhere in this chapter, the test represents only a proxy of total learning, much of which is not testable by current testing set ups (that is, standardized, large scale bubble tests). When curriculum becomes narrowed to the point of what is tested, harm occurs to students, schools, and society.

In their precient position paper, AERA (2000) warned that "curriculum and instruction may be severely distorted if high test scores per se, rather than learning, become the overriding goal of classroom instruction." Instead of being a tool of assessment, the tests have "corrupted" the very processes they were designed to assess. Critics of the testing backlash often counter argument with the claim that if the test measures reading, teach kids to read; if it measures math, teach math. They are dismissive of the entire teaching to the test argument, twisting it to insist such an approach is actually a good thing (Lazear, 2006). However, they are—once again—oversimplifying a complex issue. A test can never replicate the entirety of a knowledge set, and teaching to the test is inherently limiting and provides—in the words of the Fordham researchers—an "illusion" of student achievement.

Koretz (2008) explains how tests are "proxies" that represent only a fraction of actual learning in the same way that a sampling of voters is used in election polls, and that, while polls project the future and tests look to the past, there are considerable similarities including the unimportance of a single item, care in selecting the sampling, and recognizing margin of error. There is no practical way to survey all voters before an election, and there is no practical way to assess all the components of, say reading at a particular grade level; hence, the use of these proxies. The TAKS reading exam tests only 14% of the Texas Essential Knowledge and Skills (TEKS) that it was written to assess (Kilgo, 2006), demonstrating Koretz's argument. The problem is that 14% begins to be viewed as the domain of reading, not just a sampling of that domain. What happens to the other 86% is the great shame of reading assessment in Texas. Because the majority of the reading curriculum does not fit the parameters of TAKS (namely a large scale, inexpensive assessment), it is downplayed or ignored in favor of the 14%, that Koretz states is important "only in what it represents in the larger domain" (p. 20).

Test scores become inflated because instruction focuses only on the representation of the domain—that 14% of reading TEKS that are testable—and not the rich educational experience all students deserve. What is tested becomes what is valued; what is not valued, is not tested. It almost has a chicken-or-egg sensibility about it. Popham (2009) writes that, "[u]nassessed curriculuar aims are doomed to be overlooked" (p. 78). Such elements of the curriculum are bypassed to make room for the tested and the preparation for the tests.

2.4.4.4 Harm

Instructional time is devoted to test-taking strategies, including practice exams of released TAKS versions. Strategies in themselves are not a bad thing, but schools cross the line when they become more than just a series of lessons the week before the exam and morph into their own curriculum that can emcompass months. (I visited a classroom during *the second week of school* and heard a teacher telling the students how to set up their TAKS essays.) Popham's example is very real: the drive is not to educate, but to gain points on a test. Oftentimes, any increase in scores, and especially in schools that have been previously low performing, can be attributed not to increased learning by students, but to test prep instruction and activities (Koretz, 2008). Furthermore, Koretz points out understatedly that people who dismiss claims of test inflation as a sign the system is working are "just wrong" (Koretz, 2008).

Seife (2010) asserts that a "bad" test is one that the results of which cannot be confirmed with a different test. The Fordham researchers compared state tests to the MAP, but more often, researchers compare them to the National Assessment of Educational Progress (NAEP), aka, The Nation's Report card. NAEP scores are, as a rule, far lower than the state scores, as much as 50% in some cases. Wolk (2011) points out that, in general, NAEP proficiency rates are at about 25%, whereas some states boast rates as high as 75% for the same populations.

What are the results of adhering to teaching to the test? It produces students who can pass the state test but who cannot demonstrate those same standards/skills in class

or on "anything that is not in a test format." The test has corrupted the system and become, not the means to an end, but "the end in itself" (Wolk, 2011, p. 46.)

The myth of objectivity is also tackled by Wolk (2011), who asserts that there is little that is objective about standardized scores. Yes, they are scored by machine, but the questions were produced and selected by humans, the data will be interpreted by humans, and open-ended responses and those specifically testing a student's writing ability will be entirely graded by human standards. Farley (2009) has written very revealing accounts of what actually happens in those writing scoring sessions in *Making the Grade.* Farley writes of statistical manipulations, pressure to meet predetermined score quotas, and unqualified scorers, among other factors. Wolk is blunt on the subject: "Standardized test scores should never be used as the basis for policies that affect individual students or schools or for the formation of important public policy—period" (Wolk, 2011, p. 46).

Nichols and Berliner (2007) also put forth the argument that, "every state in America is in violation of our professional association's [AERA] norms [for responsible testing programs] because no state satisfies all of the conditions recommended by AERA (2007). Krashen echoed this concern in a posting where he claimed that, "testing students solely to evaluate their teachers" would never get through an IRB, yet Charlotte-Mecklenberg (2011) is administering 52 such tests in all subject areas strictly to evaluate teachers.

Even more troubling, is what Nichols and Berliner (2007) call collateral damage, the "corrupting [of] educators and harming [of] our common schools" (p. viii). AERA's (2000) position statement was issued well in advance of NCLB, but warned "if high-stakes testing programs are implemented in circumstances where educational resources are inadequate or where tests lack sufficient reliability and validity for their intended purposes, there is potential for serious harm" (AERA Position Statement on High-Stakes Testing in Pre-K-12 Education) AERA goes on to caution that, "students may be placed at increased risk of educational failure and dropping out" and that teachers "may be blamed or punished for inequitable resources over which they have no control" (AERA Position Statement on High-Stakes Testing in Pre-K-12 Education) both concerns were well founded.

2.4.4.5 Limitations

Interestingly, the Technical Guide for TAKS (2007-2008) repeatedly warns in chapter 10 against using TAKS as the sole source of data for making any sort of major decision or evaluation regarding students or programs (Texas Education Agency, 2007) A careful reading of the document reveals statements that seem fully at odds with the way TAKS is being used. For example, the guide reads "[t]he scale score can be used to determine whether a student achieved Commended Performance or attained Met Standard, *but it cannot be used to evaluate a student's progress across grades or subject areas*"1 (Texas Education Agency, 2009, p .85). But is that not precisely what

[1] Italics used for emphasis, not in original quote.

the public thinks the test score does show? The Technical Guide also states that, "[s]tudent test scores also may be used *in conjunction with other performance indicators* to assist campuses in making placement decisions, such as whether a student should take a reading improvement course, be placed in a gifted and talented program, or exit a bilingual program. All placement and instructional decisions *should incorporate as much data about a student as possibl*e"[2] (Texas Education Agency, 2009, p. 85). Sadly, these caveats do not seem to be widely known; students are placed in courses or tutoring groups based on this one data source alone. Programs and teachers are evaluated on the basis of these scores alone, yet the guide cautions directly to avoid such moves:

> [B]ecause the tests measure a finite set of skills with a limited set of item types, any generalizations about student achievement that are derived solely from a particular test should be made cautiously and with full reference to the fact that the conclusions were based only on that test. All evaluations of the quality of campus or district instructional programs should incorporate as much data as possible. (Texas Education Agency, 2009, pp. 87-88)

It was the purpose of this study to determine the extent of several of these practices. Why is this document not at the forefront of testing practices in Texas? Because, as Koretz (2008) asserted, this is inherently complex and what is desired is the simple. The

[2] Italics used for emphasis, not in original quote.

.

more one knows about tesing, the further it is removed from the "simple" category. There are many, many, many considerations and caveats with each test situation.

While it is understandable, but not acceptable, for teachers and administrators to think this way, it is downright wrong for those who do have the mathematical understandings to present them in such a way as to deter a wide-scale understanding of these tenets on the notion that it "keeps the commoners out" (Popham, 2009, p. 127). Navigating the literature and discussions on test construction, validity, reliability, cut scores, scale scores, and the like, is intimidating, but Popham maintains that it is "common sense" and not "esoteric mathematics" that educators need to make valid inferences from testing data (Popham, 2009, p. 127). It is worth the effort for educators to learn the basics of educational assessments.

Testing should be no more than a part of the assessment picture. In a commentary in *Education Week*, Wolk (2009) wrote of five "faulty assumptions" that continue a legacy of risk in education. One was the idea that testing is "*the* valid measure of learning" (p. 30). Overdependence on test scores is an impediment to progress and to the goals of education itself. Popham's (2009) book *Unlearned Lessons: Six Stumbling Blocks to Our Schools' Success* asserts that our current system of testing is flawed because it does not include affective assessment and is not calibrated to actually measure student learning. Popham's contention is that these tests are not "instructionally sensitive," that is, that the performance of the student does not "accurately represent the quality of instruction" the student received (p. 104). Our

automatic assumption is that these tests do reflect the quality of instruction; Popham claims nothing could be further from the truth.

Earlier, I mentioned *The Proficiency Illusion*'s findings about 7th grade TAKS reading. During the first several years of TAKS, there was a significant dip in scores from 6th to 7th grade. The reason for this was not made apparent by TEA, and caused many teachers and students across Texas great distress. I worked at a campus where the principal looked at test scores from previous years and concluded that the 7th grade reading teachers were not doing their jobs because 7th grade scores were consistently much lower than the other two grades. The principal remarked that the damage did not appear to be permanent since students rebounded to score higher in 8th grade. When I challenged the principal's interpretation, she told me "numbers don't lie." The 7th grade teachers, who were very effective teachers by every other metric, were made to feel that they had failed and were targeted for teaching interventions. This principal was drawing a common but faulty inference that has caused 7th grade reading teachers across the State of Texas much grief. The problem was not the teachers, it was not 7th graders, and it was not curriculum: It was a faulty test design.

In their 2007 report, *The Proficiency Illusion*, the Thomas B. Fordham Foundation analyzed the accountability tests from all 50 states and found that levels of proficiency varied greatly from state-to-state. Of particular interest to those 7th grade teachers in Texas, the Fordham researchers found that TAKS reading tests were not properly calibrated, notably at the 7th grade level which was proportionally harder than

any other year (Thomas Fordham Institute, 2007). The principal at my school, like many others, was drawing an incorrect inference based on what she thought was unimpeachable data; it is just not that easy.

A similar situation occurred in 2010 when the state addressed the test calibration issue by more closely aligning middle school reading tests. Seventh-grade scores showed a gain while 6th grade and 8th grade scores showed a decline. On campuses, this was interpreted to mean those teachers (6th and 8th) had not met expectations. When I visited with that same principal over scores, she noted that "sixth grade had a little trouble." This was evidence that that principal had drawn and was working with an erroneous conclusion. Once again, the complexity of the situation was camoflauged by deceptive numbers.

On another campus, a 6th grade teacher thanked me profusely for sharing this re-calibration information and the data to back it up. She said she had cried over the test scores and had been wracking her brain to see what she "had done wrong." Who seeks to measure this harm? What reformer stops to ask if data is sending a wrong message? Popham (2010) calls this stumbling block the "abysmal assessment literacy" of educators (p. 125). Who holds the policymakers, the pundits, the administrators, the journalists, etc. accountable for correctly interpreting data? And what is the harm when we do not?

Why must teachers, students, and schools be judged—and so publicly—on one measure? The National School Board Association has an idea for determining the

quality of schools called *Good Measures for Good Schools (*Azzam, 2007). They list 28 questions (or categories) to determine quality: test scores are only one. And yet, it is that one criterion, often used out of context, that wields so much power.

2.4.5 The Second Way: Why Schools Are Not—And Have Never Been—Analogous to Businesses

The notion that schools should be run like businesses is endemic in any talk of reform, but this idea has never been a good fit. In the first place, the business model may be an indicator of efficiency, but important issues remain unaddressed. For example, one aspect of the business-model reformers want to add into education is competition because they claim competition improves product. But competition produces winners and losers: some companies make lots of money (the measure of success) and others go out of business. In the overall picture, this is how the system works: the losers are those who lost money and/or jobs, but there is always (in theory) other money to be made/jobs to be had. No one is limited to one business or one job; failure is just a chance to start over.

In education the winners are schools with higher test scores (the reformers' measure of success) and the losers are those with lower test scores (which will always be high-poverty schools; see the section in this chapter on Test Score Limitations). The loser schools are targeted with punitive actions, and student losers are offered school choice (in theory, there are always available quality places for students to transfer).

Unlike businesses, failures at schools are not an opportunity to try again. The baggage of being a "failed" school or an "over age" student is heavy, and the losses are not easily recapped. Students have lost critical time that cannot be regained; the very experience of failure (and the prescribed remedies for school failure such as loss of electives, summer school, repairing grades, etc.) scars students and must be overcome to achieve progress. Educators associated with failing schools are also stigmatized and suffer professional—and sometimes personal—damage.

Rigoberto Ruelas was one of the educators impacted by *The Los Angeles Times*' database of teacher effectiveness. He taught in a high-poverty school and was well respected by his students, colleagues, and supervisors. His principal told how Ruelas had dedicated himself to his students and his community, going far beyond the expectations of his job. Still, based on scores alone, the newspaper labeled him an ineffective teacher. He took it hard, as might have been expected for someone who had his life's work labeled as "not good enough." Shortly after the results were made public, Ruelas committed suicide *(*Zavis & Barboza, 2010). Ruelas's resulting suicide is a symbol of a system run amuck.

Aside from such extreme cases, all teachers and campuses that are singled out for a perceived lack of performance carry this sort of stigma. The 6th grade teachers I referenced earlier in this section sat through a faculty presentation of slide after slide showing graphs indicating the improvements made in all other grades and content areas. Everyone knew that the 6th grade reading scores were down. Down, but not because of

anything the teachers had or had not done; down because TEA was recalibrating the test. There was no subsequent presentation pulling the faculty back together for a revised accounting of 6th grade performance that mentioned the recalibration. Under the business model, this type of activity would be meant to create competition to inspire those 6th grade teachers to work harder, but that is not the message received by either teachers or students in these settings. Schools are not businesses, nor are teachers low-level piece-work employees for whom that sort of motivation is targeted, and students are not impersonal products.

The opposite situations occur as well. Sometimes gains are merely the cause of a recalibration, a different test form, or a lower cut score. In those cases, schools and teachers are rewarded for something they did not do. The success is illusory, and the next year, those same schools and teachers will wonder what they did wrong when scores dip, as happened in New York City (this incidence will be discussed shortly).

One of the occurrences that business model proponents ignore is how Campbell's Law plays out in the business world. Berliner and Nichols (2007) call it "ubiquitous" (p. 27). When high-stakes (monetary) are attached to job performance, boundaries are crossed and what is ethical, moral, or right is sacrificed. All that matters is the deal, the sale, the quota. Consider these examples from Nichols and Berliner:

- When Sears auto repair in California began paying mechanics based on commission rather than on a straight salary, the number of unnecessary repairs skyrocketed and the state had to shut them down. The mechanics'

salaries were determined on how many repairs they sold, not how many were needed.

- Enron kept stock prices high by cheating on multiple levels until the house of cards collapsed.
- Police in a certain city were ordered to reduce the crime rate. They did this by persuading those they did arrest to confess to unsolved crimes in order to get a lighter sentence on the crimes they did commit. On paper, the crime rate went down and the department was commended—but it was all "smoke and mirrors. (Nichols & Berliner, 2007, p. 28)

Athletes in many sports, including the Olympics and the Tour de France, have been caught cheating through illegal substances because the glory (profits) outweighs the risk of being caught. Earlier in this chapter, I referenced China's endemic cheating. During the 2008 Olympics in Beijing, there was rampant speculation—and some hard data—about the age of a gymnast on the Chinese women's team (Longman & Macur, 2010). While there have long been rumors about six-time Tour de France winner Lance Armstrong's drug use, it appears that in early 2011, those rumors became substantiated (SI.com, 2011).

An example that postdates the publication of Berliner and Nichols's work, but which is certainly applicable to their example is the financial crisis which began in late 2007 with the sub-prime mortgage collapse. The pursuit of (great) profits is what drives business, even when the untold misery results.

Furthermore, Pink's (2009) work in the book *Drive* shows what motivates people to perform their jobs at high levels. It is not what one might think. Among other things, his work shows that proponents of paying teachers according to test scores are off the mark. The concept of better pay for better results has limited success even in the business world. Whenever work becomes cognitive versus mechanical, quality goes down in the presence of reward (Pink, 2009). This begs the question, do we view teachers as mechanical or cognitive workers?

Education should not operate on a business model. Nichols and Berliner (2007) propose that teachers are more like doctors, and "less corruptible than athletes and business people" (p. 29). Both educators and doctors, Berliner and Nichols maintain are critical to society and should adhere to what Hargreaves and Shirley (2009) state is "[a] compelling and inclusive moral purpose" (p. 76). In medicine, when doctors are paid by the number of patients they see rather than the quality of care they deliver, the indicator of success is corrupted and health care is compromised (Berliner & Nichols, 2007 p. 29). The indicator of a physician's success cannot be how many patients the physician manages to see in a given time period or even how many patients he or she "cures"—the indicator must be the satisfaction of the patient in the quality of the care provided. Many of the elements of that satisfaction are intangible and not quantifiable. Trying to quantify the effective doctor leads to a corruption of the service the physician provides (Berliner & Nichols, 2007).

Imagine if the mortality rates of a doctor's patients were considered the indicator of a doctor's effectiveness. Podiatrists would be near the top and oncologists near the bottom. Does that mean they are better doctors? Of course not. Medicine is too complex to use such measures for evaluation. If hospitals were also evaluated in such a manner, and hospital funding were determined on how many people survived their time in a hospital, hospitals would then choose to not admit terminal patients as a matter of their own survival, and the most vulnerable would be shut out of health care at the time they need it the most.

Those are exactly the type of indicators that fuel business models of education. Teachers whose students score high on tests are considered efficient, successful teachers, but some teachers are, as compared to doctors, "podiatrists"—their students have few academic deficiencies when they enter the classroom and minimal effort will suffice. Other teachers are "oncologists"—their students have massive risk-factors or limited English, or any number of factors that greatly impact student learning. Their "hospitals" are storefront clinics rather than advanced trauma centers. We can carry the metaphor further because it is more apropos than those that reference factory models. Teachers and students are people, not producers and products. Any indicator that ignores the complexity of the educational landscape is destined to give simplistic and inaccurate results.

Hargreaves and Shirley (2009) point out that every teacher and administrator knows that they have "good years and bad years" when factors beyond a teacher or

principal's control have deep impact upon schools. Death, illness, divorce, relocation, unemployment, and other factors significantly affect students and teachers. An inspiring principal leaves and an underwhelming successor follows. Class composition—the inclusion of a single chronically disruptive student can significantly impact the learning of all--matters. Figuring out the intricacies of a new mandated curriculum initiative can lead to a dip in students' test scores, only to have them bounce back in subsequent years when the teacher masters the material. Ditto for a shift in assessments, such as TAKS to STAAR.

Data alone cannot and does not tell the story. Even when schools do succeed according to the test score yardstick, it is important to look behind the numbers for the stories that illuminate. Hargreaves and Shirley (2009) tell of one such school that was labeled "underperforming" the year the principal's wife died, and then rebounded the next year when "he threw himself" into his work. The same applies to any student. Hardship outside of the classroom impacts achievement in the classroom. Divorce, death, a move, or any change in family structure can account for a dip in performance, yet is noted nowhere beyond the understanding of the school staff. In today's environment, data is rarely presented as just a part of a total assessment package. Washington (or Austin) cares nothing for the principal's loss, the teacher's hardship, or the student's turmoil. Their only concern: the numbers that comprise the bottom line.

In the final chapter of the now-classic *Cult of Efficiency*, Callahan (1962) calls the first era of schools-as-businesses "an American Tragedy" for four reasons. Callahan

is referring to the time period in the second decade of the 20th century, but his words indicate a continuing cycle when we consider our progress so far in the second decade of the 21st century. Callahan writes the following:

> [T]he tragedy itself was fourfold: that educational questions were subordinated to business considerations; that administrators were produced who were not, an any true sense, educators; that a scientific label was put on some very unscientific and dubious methods and practices; and that an anti-intellectual climate, already prevalent, was strengthened. (p. 246)

Callahan is writing about the period Newkirk (2009) described so well in *Holding on to Good Ideas in a Time of Bad Ones*. (Indeed, Callahan was a major source). That time period was in the early part of the 20th century, yet the leaders who were trained in that period had far-reaching influence into the 1960s. Callahan wrote in 1962 about an event in 1960 that demonstrates the end product of valuing business and industry in education over instruction and thinking. Callahan writes—chillingly to any literate person—of a censorship issue that arose in Miami just two years prior.

> Considering the professional training that administrators have been receiving in the last four decades, is it so surprising that leading school administrators in Miami, Florida, had not even heard of, much less read, George Orwell's *1984* and Aldous Huxley's *Brave New World*? This fact came to light when a parent complained because such "filthy books" were required reading in a senior English class. After reading the books, the administrators decided to bar them

> from the Miami high school curriculum. *The Commonwealth,* in reporting the incident, felt the "final turn of the screw" occurred when the U.S. Commissioner of Education, questioned about the affair, was reported to have answered, "I've never heard of those books, and I don't think it would be prudent of me to discuss them. (p. 254)

Businessmen are not educators. Schools are not businesses. Students are not products. Data is not scholarship.

2.4.6 The Second Way: When Data Collide

As indicated, data is not as clear cut as it is assumed to be. Numbers may not lie, but they can certainly be manipulated to create a lie. In his book *Proofiness: The Dark Arts of Mathematical Deception*, mathematician and journalist Seife (2010) exposes how mathematics is used to justify any number of deceptions. Seife writes that people will believe anything when a number is attached, even when that number is pulled from thin air because numbers—and this includes numbers presented in graphs—have an "aura of perfection" that is "nothing but an illusion." All numbers, Seife maintains, are flawed because they are "imperfect measurements" and are easily manipulated out of context (p. 11).

Most of those making the case for data-driven approaches do it in a most logical fashion, notably Bambrick-Santoyo (2010) in his book *Driven by Data*. However, the word "data" has been hijacked to mean only test scores from the state-mandated assessment as required in NCLB. Other sources of data—other tests, other measures,

other forms of data including qualitative measures—are not even allowed at the data-driven discussion, or if they are, they are considered lesser measures and overruled instead of being the impetus to ask more questions and dig deeper into all metrics available for use. These and other data sources are trumped by the supposed infallibility of state assessments. Because the results—including the inferences made—have not only the apparent mysteries of complex mathematical operations to bolster their acceptance, but also *scientific* approaches at their source, they are accepted without argument (Bracey, 2006). That is precisely the attitude that leads to educators being misled, or as Bracey puts it, "snookered." Bracey offers 32 non-mysterious principles of data interpretation that range from "do the arithmetic" to "beware of simple explanation to complex phenomena" to "rising test scores do not necessarily mean rising achievement" (Bracey, 2006, p. xx).

Since NCLB there has been *one* measurement for school success: scores on state standardized tests. There has been much attention on the administration of the tests, but not as much attention on the tests themselves. Seife (2010) asserts that "bad measurements" convince people of an untruth (p. 12). He gives two warning signs of bad measurements. The first is that the measurement claims to measure something that is "ill-defined," and Seife gives the example that intelligence is exceptionally hard to define and measure, "but that does not stop people from trying" (p. 12). It has not stopped Second Way thinkers from their task either, but the measure of a school is much more than student test scores. The quality of a teacher is more than student test

scores. Learning is more than the measure of student test scores. Again, remember how Koretz (2009) explains that any test is but a "proxy" for the total pool of knowledge. The proxy represents the whole; it cannot replicate it. A single test cannot ever be the measure of learning.

The second criterion of a bad assessment is that there is no other way to measure the subject and get the same results. Seife (2010) contends that when we measure the same pencil with two different rulers, the pencil is still the same length. Even if one ruler is metric, we can easily convert the numbers to get a match with our measurement in inches. However, is the same true of our NCLB state mandated tests? If our measurement is sound, the results should be duplicated with other measures. For example, in Texas, TAKS results should be duplicated on other tests, but they are not. Degrees of proficiency on the state exams are not confirmed by NAEP data, which consistently finds that states give themselves more credit for success than the national data indicates (Hacker, 2011). The Fordham Foundation (2007) in *The Proficiency Illusion* reached the same conclusion.

The inherent danger, Seife (2010) maintains, comes from our willingness to believe. "In skillful hands, phony data, bogus statistics, and bad mathematics can make the most fanciful idea, the most outrageous falsehood seem true; they can be used to bludgeon enemies, to destroy critics and to squelch debate" (Seife, 2010, p. 3). Seife claims these actions occur in all facets of our lives and that this "proofiness" (which he defines as "the art of using bogus mathematical arguments to prove something that you

know in your heart is true—even when it's not.") is a serious threat: "In short, bad math is undermining our democracy" (Seife, 2010, p. 5)

Bad math has certainly undermined education. It is continually used in all three of the situations that Seife (2010) lists, but it is also used to establish need where none exists, cover up the performance of programs, demonstrate success (or lack thereof) when success (or lack of) is not present, and as a powerful sales tool. Rod Paige, former superintendent of Houston ISD and former Secretary of Education, was given credit for the "Texas Miracle" where Houston allegedly showed a dramatic decrease in the dropout rate by raising standards and instituting financial incentives for principals at successful schools (and reassignments or dismissals for principals at less than successful schools). However, the "Texas Miracle has since been exposed as something more resembling the Texas Fraud" (Robbins, 2007, p. 88). A *60 Minutes* investigative report showed that while Houston ISD reported a dropout rate of 1.5%, the actual rate was somewhere between 25% and 50%. They managed to hide their dropouts by recoding students (at the direction of Superintendent Paige) to reduce the numbers (*60 Minutes*, 2004). Furthermore, the high test scores achieved were also manipulated. Again, under pressure to raise scores or depart, principals made sure some students did not take the tests by such measures as having students skip tested grades, retaining students in the grade prior to a tested grade, or forcing students to drop out before testing (*60 Minutes, 2004*). And Houston was "both the impetus and model for NCLB" (Robbins, 2007, p. 88).

The promise of NCLB was that by identifying students who were underperforming through data, schools could close the "achievement gap." Previously, the gap had been hidden in campus or district scores that reported an entire population. Disaggregating the data by poverty and population groups exposed these gaps and forced schools to, in the words of the legislation, "leave no child behind" since every group counted and could not be combined with a higher achieving group to disguise the low performance. Early on, the legislation had its critics who believed the law was overly simplistic and ill-funded (Fusarelli, 2004). Certainly, the Texas program that NCLB was patterned on had its critics as well (including Berliner), but other researchers found particular fault with one of the population groups NCLB was supposed to be helping: English learners.

During the TAKS era in Texas, researcher MacNeil (2000) found that the policies meant to promote student achievement were actually having the opposite effect. MacNeil quotes the militaristic verbiage used by Ross Perot, who was appointed by then Governor Mark White to oversee the Select Committee on Public Education (SCOPE), who said that, "[w]e've got to nuke this educational system," and points out that the "bombs" fell not where Perot intended. His belief was that teachers needed help in the form of rescue from district level administration and building administrators. Perot had a strong antipathy towards administrators, especially those who had been coaches. But they were not the casualties in this rhetorical battle. MacNeil points out that these bombs "fell instead on classrooms, on the teachers, and on their students"

(MacNeil, 2000, p. 189). MacNeil also uses the term Berliner does: "collateral damage." These reforms led to "distrust and a cycle of lowered expectations" that still exist in Texas. Standardization of assessment, MacNeil argues, does not raise the low achievers, but rather, "compromise[s] higher academic quality" (p. 215). It only indicates that these assessments cause harm to *all* students.

Popham (2009), respected researcher, author, and professor emeritus at UCLA, posits that a major issue hampering educational reform is the "abysmal assessment literacy of educators" (p. 125). The data-driven mandate that was thrust upon them included no provision for understanding the data that was supposedly driving their every move. Because the "assessment literacy" is so critical, Bracey's book and Koretz's *Measuring Up* are perfect antidotes to the "abysmal state" Popham describes. Koretz (2008) argues that everyone wants simple answers to complex questions—even if they are wrong. Politicians, some reformers, parents, realtors, and even some educators want a measure that simply and conclusively shows if children are being taught effectively. They want "a simple explanation" of the "complex phenomena" that is education as it relates to the individual student, the teacher, the campus, the district, and the larger community (both state and federal; Bracey, 2006, p. 35). Again, it is just not that easy. Education is a very complex endeavor; the assessment of education is even more so, and far more so than any single metric can describe.

Chapter 1 opened with a photograph of an infamous sign outside the town of New Cuyama, California. The arithmetic is correct, but the numbers do lie because

these are three numbers that were never meant to be added together, and any conclusion drawn from this total is bound to be in error because the person who did the addition did not understand the purpose of those three numbers above the line. (In the case of this photo, it is obviously a joke, but this type of erroneous conclusion is not restricted to humor.) Such is often the case with educational testing data. Frequent misinterpretations involve comparing different groups and claiming either growth or decline (such as "this year's 7th graders scored higher than last year's 7th graders," which is not valid for TAKS because it features different students on different assessments), or making a claim like "our growth in math was double that of reading" (Bracey, 2006, p. 31). Interestingly enough, the TAKS Technical Digest—in a great moment of doublespeak after saying such comparisions are not valid—says these comparisons can be made, but provides a very important caveat I will discuss below (Texas Education Agency, 2009).

Assuming causality (that is, claiming that one element causes another) is also a frequent mistake. For example, success in reading may be attributed to an after-school tutoring session for those at risk, but improvements in general instruction, the amount of time spent reading/number of books read in a year, fluctuations in cut scores and pass rates on the test are not even considered as having a part in the causality. What percent of the gain is owned by each of these potential causes? Did any of them have a negative effect? Those are important questions that are never addressed. We see in data what it is we want to see.

Again, this is a case of oversimplyfing a very complex issue. But perhaps the most common—and most logical sounding—is the desire to have students achieving at grade level. This belief is so common that it even has a name: the Lake Wobegon Effect, after Garrison Keiler's mythical town where "all the children are above average" (Koretz, 2008, p. 58). How many parents, policymakers, and even educators realize that grade level is simply the median point of a set of data? There is no standard list of skills that constitute grade-level achievement, and it is, frankly, impossible to have all children score above the median (Allington, 2008).

Educators are encouraged to "drill down" into the data to draw conclusions. Popham (2008) calls this practice "educational silliness," and says that teachers do not have time for such dead-end activities. A good example of this would be the 2009-2010 9^{th} grade math TAKS. There are 52 questions, representing 46 different student expectations or SEs (TEA, released TAKS key). One SE has four questions, another has two, and the rest have one question apiece. Only the SE with four questions can be considered to have had enough questions to truly be assessing a skill. The others do not have sufficient test items to make a valid inference (Popham, 2003). Reading presents a different situation, as cited earlier. Only 14% of reading SEs are tested on TAKS and the other 86% are considered untestable. The data tracking system used can produce pages upon pages of data to indicate which population groups got the question correct and which groups did not. Teachers are taught to look at any question with more than a 10% error rate, and consider that an area that calls for reteaching and/or tutoring

(Crook, 2005; Kilgo, 2007). This is an example of using summative data as diagnostic, or formative data.

In order to get valid data on whether a student truly understands a concept (SE) or not, multiple questions are needed—several more than the one question per SE the math test utilizes (Popham, 2003). The teacher cannot be sure of why the student missed the question. Was it a lack of comprehension in reading the problem since all TAKS math questions are story problems? Did the student transpose numbers? Add or subtract incorrectly in a multi-step problem? Bubble in the wrong answer choice? There are potentially many reasons for the error, but only one conclusion is drawn: the student does not know the SE. Without seeing the test question and without talking to the student about his thought process in working out the problem, any conclusion has the potential to be as erroneous as that New Cuyama, California sign.

Another example of drilling down to "silliness" comes on that reading portion of the TAKS. Reading is different in that the test measures the successful application of skills, not the understanding of skills themselves. Questions are tied to individual SEs on concepts like main idea, context clues, inferences, etc. Applying the Crook and Kilgo methods (as many Texas schools do) to TAKS data, teachers reach the conclusion that their students are weak in locating the main idea, determining context clues or the like. There may be four main idea questions on a TAKS test—a generous number really, since many SEs have only one or two questions—and if a student misses two of them, it is believed that reteaching the main idea is needed, either in a whole class

setting or in tutoring sessions; but, is that really the case? The test did not ask what main idea was, it asked students to determine the main idea of a particular passage. Could the student have understood the concept of main idea and still missed the question? Absolutely—remember, two of the four questions were correct. What the student failed to do was identify the main idea in the particular passage. Why? There are reasons other than a lack of main idea understanding that could account for the scores. Perhaps, the student did not connect with the passage or was confused by vocabulary in the passage or answer choices. Perhaps the question came at the end of the test when the student was tired and just randomly filling in bubbles. Perhaps the student was relying too much on prior knowledge instead of what was on the page. As it is with math, without the test and without talking to the student about what he or she is thinking, it is impossible to know why the question was missed.

The harm in each of these examples is that, to remedy these percieved deficiencies, instruction is compromised. In too many reading classes across the state, reading rich and challenging text is replaced with TAKS-like passages that are mined for these basic skill concepts. I might add that middle school TAKS passages are written by the test makers; high school passages must come from published materials (TAKS Blueprint, 2007). The high school materials are of a better quality, whereas the the middle school passages are pale imitations of the high quality texts that could populate a middle school reading class. The steady diet of test prep is becoming more

and more entrenched and taking up more and more instructional time (Harris et al., 2011, p. 96).

Each of the researchers Popham, Koretz, and Marzano, has addressed how the lack of a solid understanding of assessment data gives educators a misguided sense of what testing can do. Reeves (2008) argues that it is time for us now to recognize the limitations of quantitative data and pull in more qualitative data to flesh out the assessment picture. Reeves expresses that we must begin to look for "the stories behind the data" (p. 89). However, it is unlikely that will take place on a large-scale level such as NCLB. Qualitative measures are more expensive than bubble tests, and do not lend themselves to the same level of standardization. Additionally, testing is already quite pricey in terms of development and administration. Testing is now a $1.1 billion a year industry; Texas is paying Pearson nearly half a billion dollars over five years for the new STAAR (Smith, 2011, Vu, 2008). However, qualitative measures do more to reflect the reality of the situation—that teachers and students and learning are not standardized.

Popham (2009) writes that educators' lack of assessment fluency comes from a fear of math. There is much that can go wrong when interpreting data. Seife (2010) lays out three ways people lie with numbers: Potemkin numbers, disestimation, and fruit-packing, which includes apple-polishing, cherry-picking, and comparing apples to oranges. Seife then goes further to explain three ways that people misinterpret numbers:

"randumbness," "causuistry," and "regression to the moon" (pp. 55-66). All of these are present in today's data-driven educational culture in our schools.

Seife (2010) writes that the term "Potemkin numbers" comes from a legend about Prince Grigory Potemkin's deception of Empress Catherine the Great of Russia. He wanted her to think an unpopulated area she traveled through was, in fact, a thriving region. Potemkin erected building fronts far enough from the traveling route, which—at a distance—gave the appearance of villages and towns. Behind the façade was *nothing.* Potemkin numbers were numbers that were phony but appeared to be real. They can be totally ungrounded such as an advertisement's claim of their product producing 23 times more protection (how do you measure the protection of, say, mouthwash?) or have some basis in reality. Seife explains how, in a similar case of the total attendance at the Million Man March held in Washington DC in 1995, various attendance figures for that event were computed and disputed, and how the political fallout stopped the Park Service from estimating crowd size for ten years, allowing organizers to create their own Potemkin numbers.

In education, the blatant attempt to deceive is not usually present, but perception of certain numbers creates deceptive conclusions. For example, grade-level proficiency is cousin to a Potemkin number. It sounds logical and admirable, but once we glance behind the façade we realize that there is nothing there. In the NCLB legislation, the expressed mandate was to have every child on grade level by 2014 (Ravitch & Chubb, 2009). It sounds laudable, but it is mathematically impossible. While almost everyone

assumes that there is a yardstick for measuring grade-level achievement (i.e., things students should learn at each particular grade), none actually exists. Standards provide some guidance, but those are not consistent across states, and many are vaguely and illogically written. For example, in Texas there is a 1st grade student expectation in the TEKS that requires first graders to "identify and describe the roles of public officials in the community, state and nation." Fortunately, the SE that read "identify and describe the role of a good citizen in maintaining a constitutional republic"—which in itself is considerable—dropped the words "and in keeping elected officials responsive to the wishes of the people" (TEA, Elementary TEKS Revision, First Grade, 12C).

Allington (2008) writes that grade level is simply the point where fifty percent of students score above and fifty percent score below—a median. Work hard to improve the performance of that bottom half, and the line moves up. It is a no-win situation. There will always be a bottom half, except in Garrison Keillor's Lake Woebegon, "where all the students are above average." Perhaps NCLB refers to the grade-level equivalent scores provided by such tests as the Gates MacGinitie and Nelson Denny reading assessments, or the Iowa Tests of Basic Skills (ITBS) and other achievement tests. However, those are normed reference tests and the grade levels are determined by comparing the performance of the test taker to the test takers in the norming pool. A student who scores, say, 10.8 is not necessarily ready to move to the end of 10th grade; it means the student scored as well on the test as 10th graders in their eighth month. Helpful data, but it does not describe the academic performance level of 10th grade.

Also, the very nature of a normed test (that results will be spread out) is contrary to the goals of NCLB, which call for a more criterion-referenced format where achievement can fall into natural patterns. Additionally, these numbers will shift when the test is re-normed since the numbers come from a comparison to other students and not to a set criteria of knowledge and skills.

Reading materials frequently—and especially in school—are given a grade level equivalent to indicate the difficulty of text such as 7.5 indicating that text is at a Grade 7, 5th month level. Any text that falls within a 7.0-7.9 range would be considered to be on a Grade 7 reading level. Those numbers are computed using a formula that considers word and sentence length, and thanks to technology, have become exceptionally easy to use. For example, the Flesch-Kincaid reading levels are included in Microsoft Word and can be computed by pasting a text sample into a Word document and running the spell check function. The problem is that the computer (or formula if computing by hand) cannot take into account the subject matter of the text. Thus, Hemingway's work usually comes in around 4th-grade level, and science textbooks come in above the grade they are ostensibly written for. Trying to decipher if students are on grade level based on the level of the texts they read is not an accurate method of determining grade-level proficiency.

Grade-level proficiency is a type of Potemkin number. Potemkin numbers are intended to deceive, but in this case, the numbers are merely deceptive by interpretation. The intent is different, but the results are the same: people assume a falsehood.

Fortunately, as Seife (2010) points out, it is an easy type of proofiness to detect, requiring only a closer look. As soon as someone begins to ask questions and thoughtfully evaluates claims made, the numbers collapse. Familiarity with Hemingway shows that *A Farewell to Arms* is not an appropriate book to read aloud to 4th graders, and that although 9th graders are using the 9th grade biology book, its reading level is 10.9, so scaffolds must be utilized to help students comprehend the test. According to Pomham (2009), educators need to learn to ask the appropriate questions of data by overcoming the "abysmal assessment literacy" (p. 125).

Since there is no set definition for proficiency at a particular grade level, how can we know if a student has met or exceeded it? If we are using the grade equivalent scores from a normed test, we must realize the comparative nature of the test and look for additional evidence to either support or refute the score. These are examples of the types of thinking that can be the only way to eliminate the illusion of Potemkin numbers.

Much more dangerous than Potemkin numbers is *disestimation*, which Seife (2010) defines as "taking a number too literally" and not being skeptical enough to analyze it (p. 23). These numbers are derived from what Seife calls "a real, meaningful, good-faith measurement," but he points out that all measurements are flawed to some degree (p. 20). That error can be reduced with "more precise (and expensive)" instruments, but some degree of error is inherent (p. 20). Seife uses the example of a pencil to show that, while a simple ruler will give a respectable measurement, there are

other devices in laboratory settings that will give even more precise measurements. However, those instruments are not available to the student who needs to measure the length of his or her pencil. Neither are precise instruments for measuring student achievement in a large scale setting available to schools. When a student is tested for special education a barrage of tests are administered individually to gain a complete and precise picture of the student. However, the expense in time, personnel, and resources are not justified in spreading this type of testing to a larger audience. No one is suggesting that such an action take place, but it shows how the large-scale tests are like the ruler used to measure the pencil. Good enough, perhaps, for a measurement for some purposes, but not specific enough for others.

The way in which test scores are used is an example of disestimation. They purport to measure student achievement, but with what margin of error? TAKS and the other NCLB tests (and the SAT and ACT, for that matter) are designed to be administered to large numbers of students as inexpensively as possible and return the results as quickly as possible: large scale, fast, and cheap. Choosing such a method means sacrificing some degree of accuracy (Seife, 2010). For example, a school may decide to determine reading levels for each of its students. A large scale, fast, and relatively cheap way to get this data would be to administer the Gates-MacGinitie reading test. For about an hour of class time for students to read and bubble plus a few minutes of scanning time, a teacher or administrator can have data on the entire school population. But is the data as helpful and/or accurate as the data gained from

administering the Developmental Reading Assessment (DRA)? The DRA is not fast; it must be administered individually and takes easily 30 minutes of teacher time to administer and score for each student. Students read aloud, discuss with the teacher, read silently, answer questions about their reading and metacognitive processes, and then write. It is a much more comprehensive and differentiated assessment (students will be reading different texts according to their proficiency levels; Gates is the same text for all test takers). For the convenience of the Gates, schools sacrifice the accuracy and detail of the DRA which is just too labor intensive for a large scale administration in most schools. But because that is known, Gates data is taken—in this instance—with that proverbial grain of salt. It is what it is, but there are limiting factors that should be considered when bringing it to the table for decision making.

In TEA's TAKS manual, the agency specifically cautions against making decisions based on TAKS scores alone. The 2100 passing score (or to be technical, the score that indicates the standard has been met) does not make it a "magic" number that separates the literate from the illiterate or even show that a student is capable of doing grade-level work. The Fordham Foundation's (2007) biggest concern reported in *The Proficiency Illusion* is that millions of parents are lulled into believing their children are progressing in their development of reading skills, when, in fact, they are not progressing at all because the testing bar is too low—it is disestimation at work. Like political polls, all tests have a margin of error; the larger the scale of the test, the greater the error. However, unlike polls, NCLB does not account for margin of error. Nor does

it consider the type of thinking that reveals Potemkin numbers, but both must be applied to fight disestimation of those scores. Seife (2011) writes that the danger in disestimation is that the numbers mix fact (2100 signifies met standard) and fiction (2100 signifies that a student has grade-level proficiency) and that once rooted, takes a long time to fade away.

The final category of mathematical deception, according to Seife, is fruit-packing, which has three manifestations: cherry-picking, apple-polishing, and comparing apples to oranges. All of these are well represented in education. Cherry-picking is the focus on data that favors one conclusion and the exclusion of data that does not support that conclusion. Seife (2011) outlines the way a decision should be data-driven:

> Since real-world numbers are fuzzy, answers to numerical questions aren't always clear cut. Measuring the same thing in different ways can give different answers; some of the numbers will be too high, some will be too low, and, with luck others will be reasonably close to the right answer. The best way to figure out where the truth lies is to look at all the data together, figuring out the advantages and disadvantages of each kind of measurement so that you get as close to the truth as possible. (p. 27)

The truth is not the goal of the cherry-picker; accomplishing an agenda is. Seife (2011) calls cherry-picking "lying by exclusion to make an argument seem more compelling" (p. 27). No matter the cause, cherry-picked data is used for support. In his

book, Seife takes politicians from both sides of the aisle to task for their use of selective data, including George W. Bush, who in 2007 touted the success of NCLB by claiming that students had higher achievement in both reading and math since the passage of that legislation. Seife asserts that Bush uses two incidents of cherry-picking to make this claim. First, Bush used NAEP data from 4th and 8th grades that show a small increase in scores, but ignores the 12th-grade NAEP scores, which show a decline. Bush also used data only from 2002—when NCLB became law— to support his claim. However, math and reading scores in those grades have been rising steadily (albeit slowly) since the 1990s—long before NCLB. Bush took the tail end of a trend and claimed credit. In addition, by focusing only on math and reading, Bush has cherry-picked from school subjects (i.e., history, languages, writing, music, etc.) that do not show gain or are not tested.

Apple-polishing involves packaging the data in ways that emphasize the more favorable aspects and downplay the less favorable.. Graphic representations of data are especially vulnerable to this type of manipulation, causing Popham's concerns about the "abysmal assessment literacy" which comes into focus once more. Educators must be critical consumers of numbers and must educate themselves to understand the limitations inherent in measuring student achievement. Other forms of apple-polishing include inappropriate use of mathematical functions and operations, including the use of averages. Many numbers that ought not to be averaged are, thereby skewing the data to give a mistaken impression. Averages are particularly vulnerable to outliers, which can

either elevate or depress the score. The average income of a town where say, a Bill Gates or Warren Buffett lives, is skewed from the neighboring town with similar demographics but which lacks a billionaire resident. That is why a median number (where half the population falls on either side) is a more precise measurement than the average. Apple-polishing is choosing the best angle of data—not the most accurate or truthful—to get the desired result (Seife, 2011).

The headlines are frequently filled with the final expression of fruit-packing, comparing apples and oranges. As expressed earlier, education is profoundly situational and valid comparisons are not that easy. The landmark report, *A Nation at Risk,* took international student comparisons and became convinced the sky was falling (it was not; Bracey, 2006). Virtually, every press release for international tests has been greeted in the same manner. The 2010 PISA (Programme for International Student Assessment) results were announced in December of 2010 and immediately the hyperbole began. Arne Duncan called the results "a wake up call" (MSNBC, 2010). To what? Reform has been a major topic since *A Nation at Risk*. Schools have been in crisis during virtually every decade of the 20th (and now 21st) century (Ravitch, 2011).

The comparison of the performance of U.S. students to students in other countries is an apples-to-oranges situation for several reasons. First, cultural expectations are vastly different. Without going into too much detail, consider the life of the student in China and a student in America. The American students have far more on their plates than school. The American student has to juggle a combination of

activities such as athletics, music lessons, band, drama, clubs, jobs, social activities, familial responsibilities, faith-based activities, volunteer work, and more. The Chinese student has one: school (Reynolds, 2008). Is that really what America wants for its schoolchildren?

Our schools are—for better or worse—whole child and community focused. We have arts, athletics, and social organizations. We recruit for the military in our high schools, we prepare for both college and vocational careers, and we are inclusive of all levels of achievements. We educate special education students alongside mainstream students (and even after, up to age 21). We do not have special schools for athletes to feed our sports programs. Our schools represent our communities: urban, suburban, and rural. High school homecomings generate more interest than testing days. Class reunions and alumni events pull former students back. Booster clubs for sports, band, and drama keep parents highly involved with schools. In rural communities, the school is the lifeblood of the area. In America, schools are so much more than data points. This is not to imply that other nations do not also address these issues in ways that develop the individuality of their students, but rather to show that in America we expect our schools to deliver much more than curriculum.

Additionally, we do not have a standardized curriculum across the country as other nations do. The Department of Education was established with the specific mandate not to establish curricula, but to respect that education was the province of the state. Encroachment on that receives pushback, as is now apparent with the Common

Core Standards, a document that was created by the Coalition of State Governors but endorsed by the DOE, deftly sidestepping their mandate to stay out of the curriculum business (Mathis, 2010). So when American students take international exams, their preparation is incredibly diverse: fifty different sets of standards govern their instruction. In other countries where education is the province of the nation, there is only one curriculum and one set of standards.

Another consideration is this: maybe our kids are just tested out. What are the circumstances of test administration? Who are these kids? Where are they? There is a standing assumption in the world of education reform that tests represent a good faith effort by students. In reality, that is not the case. Students are often not interested in tests or test results, especially those that they do not see as having any relevance for them. In Texas, there is much discussion over the lower success rates on the 10th grade tests since students see those as "the one that does not count" whereas the 11th grade tests are the exit exams required for graduation. This is not a new phenomenon. I remember drawing Christmas trees and fish on my Iowa Tests of Basic Skills in elementary school because they seemed so long and I disliked the mustard color of the ink on the answer document.

Should this data ever be more than interesting? When we know so little about the full context of these exams, why then are they such cause for alarm among policymakers? Are students really so underprepared? There is evidence to suggest otherwise.

In her book *The Overachievers: The Secret Lives of Driven Kids*, journalist Robbins (2011) spent a year with overachieving high school students to report the first-hand effects of today's high-pressure, high-expectation culture on "the best and the brightest." Her work uncovered an intense experience for these students with very little quality of life or quality of education. One of the most telling comments comes late in the book, when a California senior relates, "My teacher once said someone else's success is not your failure. I wish more kids realized that, because a lot of nasty things have been said about kids and their schools" (p. 326); that is a statement worth considering with the PISA results. Hooray for Shanghai, Singapore, and Korea. Congratulations for finishing at the top. But does that mean that America is so terrible? Think about it and consider the graduation requirements for high school. Never at any time in the history of our nation have standards been so high. Students in Texas are taking 4 years of math and science (compared to the 1 year I was required to take) in addition to a world language, 4 years of English, 4 years of social studies, and various requirements regarding technology. Why can this not be celebrated for the amazing feat that it is?

Robbins (2011) states it plainly, "We live in the Age of Comparison. Too often we deem our own achievements worthless if they fall short of others' standards. Our best isn't good enough if it's not as good as someone else's best" (p. 326). Does the United States really want to have an educational system (and society—they are inseparable) like Shanghai, Singapore, or Korea? Zhao (2009) makes an emphatic case

for *no,* and writes of the strengths of the American system that are underreported. Comparing apples to oranges results in this sort of discontent and the application of faulty logic: "In order for someone to succeed in the Age of Comparison, someone has to fail" (Robbins, 2011, p. 327).

Another example of this faulty comparison is the comparison of test data from year-to-year. Again, it sounds entirely logical to make those comparisons and claim growth when scores increase. That is what the mayor of New York City, Michael Bloomberg, did. Scores on the state exams have risen noticeably since 2005. Bloomberg was quick to claim this showed the reforms he implemented were working. However, the NAEP scores during that time did not rise. Teachers noticed that the 2005 state test was much easier than the 2004 test, and it was soon revealed that the State of New York had "tinkered" with the tests, making them easier and making score comparisons invalid (Seife, 2010). Since the tests after 2005 were not as difficult as those given in 2004 and before, it is becomes an apples-to-oranges comparison. Students scored higher on an easier test, and that does not demonstrate increased knowledge.

Texas is in the same situation. Although the TAKS technical manual specifies that year-to-year comparisons for individual students (TEA, 2007, Chap. 10) cannot be legitimately made (the tests are not the same; the students are not the same), invariably the comparison to previous years is made. In 2010, TAKS results were unveiled in the spring and headlines announced they were the highest level yet. Then over the summer, an investigative reporter for the *Houston Chronicle* exposed that scores were higher this

year—an election year—because cut scores (the point at which students meet the standards, or pass) had been lowered. TEA claims they need this leeway to make accommodation when a test is "harder" than its predecessor, but this raises questions about the consistency of the assessments and the validity of claims that students are learning more.

Seife (2010) coins the word "randumbness" to indicate looking at numbers and perceiving a pattern where there is none (p. 55.) He argues seeing patterns is hard-wired into people and that it is very easy to read too much into data. Seife writes that "[i]t gets even very smart people to believe idiotic things" and cites serious research showing that athletes who wore red in certain events during the 2004 Olympic games won more matches than athletes who wore blue (pp. 55-56). An interesting note, certainly, but one whose predictive value is nil. It was a random collection of data and doomed to be refuted in any future study because the color had nothing to do with the factors that do impact athletic success. Schools frequently fall for this line of thinking. It goes like this, in my opinion:

1. We want to be an exemplary campus like Pearly Gates Middle School.
2. Pearly Gates does not use student lockers.
3. Therefore, we will stop using student lockers.
4. Then we will be an exemplary campus like Pearly Gates.

Furthermore, even if there is a pattern there is no guarantee that the pattern is relevant. In such cases there is little distinction between a pattern and superstition.

Seife (2010) also points out that drawing a line through any charted set of data, leads the observer to see a pattern even if one is not present. This line of regression can be an "extremely powerful technique," but it if there is no pattern to the data, the results are "meaningless or downright barmy" (Seife, 2010, p. 63).

2.4.7 The Second Way: The Tyranny of Data

As discussed earlier, the Technical Guide for TAKS repeatedly warns against using TAKS as the sole source of data for making any sort of major decision or evaluation (TEA, 2007). The Technical Guide (2007) also states that,

> Student test scores also may be used *in conjunction with other performance indicators* to assist campuses in making placement decisions, such as whether a student should take a reading improvement course, be placed in a gifted and talented program, or exit a bilingual program. All placement and instructional decisions *should incorporate as much data about a student as possibl*e. (p. 87)

Yet TAKS scores reign supreme as evaluators. Many reform efforts—including merit pay for teachers—rest solely on these numbers, in spite of the caveat attached to them. Why is this document (and these ideas) not at the forefront of testing practices? Because, as Koretz (2008) asserted, this is inherently complex and what is desired is the simple. Blame that and the "abysmal assessment literacy."

Berliner and Nichols (2007) add that, "the testing program we have bought into destroys motivation to excel, rewarding, instead, motivation to get by" (p. 10). This is an unexpected consequence of a system that was meant to encourage students and

teachers to achieve more, but the system was ill-conceived and shows no connection to what Pink's (2011) research shows clearly motivates both groups. The "threats and incentives" model does not produce the desired effect in education because education is too complex for simple solutions. It is not—as is depicted in the documentary *Waiting for Superman*—a simple matter of passive students having knowledge poured into their heads.

It is indeed ironic that when George W. Bush launched NCLB, he frequently mentioned the "soft bigotry of low expectations" to describe how some students are quietly yet systematically denied success. He was correct in that certain populations of students were being underserved, but the irony is that now those student populations have glaring spotlights upon them. However, the simplistic formula of attention + incentive = success is not creating success for these children. Berliner and Nichols (2007) write that, "those who fail at tests often attribute their failure to a lack of intelligence, a belief communicated to them also by schools that see those lower performing students as hurting the school's reputation because of their low ability" (p. 11). The very students who are meant to be helped by NCLB are often being singled out by means of extra tutoring sessions, special classes, and special assemblies. One administrator, Joseph Showell of Grand Prairie High School, himself African American, called all the African American "bubble" students into the auditorium. He told them it was the scores in the African American population category that were keeping the school's accountability ranking down. He hoped to motivate them to higher

achievement, but the students at the assembly felt the principal was telling them that African American students "are stupid" (as cited in Chavez, 2009, p. 1A).

Berliner and Nichols (2007) write that, "[a] new form of discrimination is apparently creeping into our schools, and it is against the score suppressors, those children who keep the school from not looking good on high-stakes tests" (p. 11). Berliner and Nichols (2007), caution that students who feel they are stupid, unlucky, or have bad teachers are not likely to make any attempt to study or perform better on the next test. Most of these students drop out of school, the researchers maintain, as a way to save face in an atmosphere where they believe they are marked.

All of this could be just dismissed as obstructionist and favoring the status quo (and often is by some education reformers who further complicate the arena with ad hominem attacks and skillful manipulation of facts), except that the evidence boils down to this statement, made by Berliner in 2007 and confirmed by an exhaustive study by the National Research Council of the National Academy of Sciences that found it is not clear that learning has improved in the testing era. (Hout & Elliot, 2011). The subtitle of *The Myths of Standardized Testing* is "Why They Don't Tell You What You Think They Do," which illustrates once more that testing does not meet our expectations. We expect standardized testing to be a scientific, accurate mode of assessing student learning, but it falls far short.

2.5 The Third Way: A Promise of Synthesis

The Third Way was meant to bring in the promise of synthesis: the best of the two previous movements and establish a true collaboration between those dictating policy and those who were implementing it. During the mid to late 1990s, there were some steps toward progress in education, an invitation for all parties to come together and share to create a product that offered the best of both First and Second Way thought (Hargreaves & Shirley, 2009).

2.5.1 The Third Way: Distractions

The short-lived Third Way was a shortcut back to full Second Way thought in education. Hargreaves (2009) identifies three "distractions" that led to this about-face, explaining how it is that we are here today. For an approach that offered such promise for true educational reform, the Third Way turned out to be anything but.

2.5.1.1 "The Path of Autocracy: Top Down Delivery"

The Third Way was meant to be the way of collaboration, of give and take, of consensus; however, it soon became apparent that those who hold power do not easily surrender it. Hargreaves and Shirley (2010) write that, "what was meant to be a strategy of development . . . has actually turned into a slickly spun system of top-down delivery" (p. 23). Yes, during the Third Way era states developed standards, assessments, and tracking systems, but it was all under the direction (threat?) of NCLB and increasing federal influence upon schools. Reports from this period continue "the sky is falling" on American education theme that began with *A Nation at Risk* and advocate more of what

Hargreaves and Shirley (2009) tag as "the unholy trinity: markets, testing and accountability." Hargreaves and Shirley are not defenders of the status quo, rather they agree with many of the criticisms of education, but strongly object to the notion that the "trinity" will solve any of the problems now facing education. They assert that such tactics are based upon not evidence, but instead "ingrained ideology, not about learning from the evidence of other people's successes elsewhere" (p. 25). Hargreaves and Shirley point out that the very nations who are beating American students in international testing would not even think of utilizing the reforms American schools are increasingly being forced to adopt presently.

Hargreaves and Shirley (2009) argue as well that these ideological devices—market, testing and accountability to testing—are perpetuating the problem by insuring that "deliverology" (defined as designing instruction to meet narrow targets on an assessment vs. a deep, rich education that builds capacity) is the main role of teachers. The language Hargreaves and Shirley use is strong, and aligned with Berliner and Nichols' (2007) concept of "collateral damage":

> Pack your building with teachers who have a single-minded focus on raising test scores in the basics and you don't have a learning organization, but rather an ingrained distraction from the core tasks of teaching and learning in a diverse community setting. (p. 26)

Furthermore, Hargreaves and Shirley (2009) argue, so called failing schools will always be doomed to failure because whenever they approach the success marker, the

government/state moves the marker, making it appear on the reports and in the media that the school is not making progress. They cite the situation in England where the bar set by the government is continually raised, undercutting the actual achievement of the schools, which "manufacture[s] the exaggerated appearances of failure that justify its [government's] control and intervention" (p. 27).

The same situation is unfolding in Texas with the end of TAKS and introduction of the (supposedly) more rigorous STAAR system with its multiple End of Course exams in high school. The last time the state switched assessment programs (from TAAS to TAKS) was before the advent of such huge public emphasis on test scores, and the shift with its decrease in scores for all schools (until the test became more familiar) went largely unnoticed by those outside the field of education. When scores go down on the new tests, how will these "baseline" scores be received? Given how the public seizes on any perceived negative news and ignores the positive (as shown elsewhere in this chapter), it has to be an area of concern for Texas school districts. Will this become a rallying cry for those who demand more control over schools, or can schools utilize this instance to demand more local control and less state/federal interference? Hargreaves and Shirley (2009) note that as these scores have historically risen with each administration, the government steps in to claim success for the rise that is, in effect, just a "recovery."

2.5.1.2 "The Path of Technocracy: More Data Driven Nonsense"

During the first decade of this new century, technological advances made the acquisition of data both easy and abundant. Twenty years earlier the manpower needed to produce something as basic as an item analysis of a test without a computer was not feasible for most school districts. Now it is available in seconds. Because we have access now to data in any conceivable configuration, Hargreaves and Shirley (2009) assert that we have become too enamored of our Frankenstein. Just because data can be produced by these technologies does not mean it is relevant, insightful, or even true, but in our zeal to become "data-driven," we have ignored those issues. The Path of Technocracy, Hargreaves and Shirley claim has "converted moral issues of inequality and social justice that should be a shared social responsibility into technical calculations of student progress targets and achievement gaps that exist around the world" and made them apply to only one environment: the school (p. 29). Schools alone are blamed for these gaps, and to remedy them, even more data is collected.

> Moral issues and responsibilities are converted into technical issues and responsibilities to be resolved through ever-increasing testing and analysis of voluminous amounts of numerical achievement data. Evidence is not only collected about people, but individuals are also increasingly required to collect data about themselves and each other to check and chart their progress on a never ending basis. (Hargreaves & Shirley, 2009, p. 29)

In the proper context, they argue, data can "inform" decisions, provide clarity, generate questions, spur action, challenge our perceptions, etc. But the Path of Technocracy does not acknowledge data in that fashion, instead offering "exorbitant promises about what data can do" while ignoring that "[p]rofessional judgment and experience" are needed to effectively understand and utilize data (Hargreaves & Shirley, 2009, p. 31). This issue was explored in sections 2.4.6 and 2.4.7 of this chapter in greater detail, but it is important to note that this Third Way distraction is what has firmly rooted education in that Second Way approach. "Technocrats value what they measure, instead of measuring what they value," and this mindset destroys the promise of the Third Way (Hargreaves & Shirley, 2009, p. 31). Personal factors (death, new curricula, "the difficult class") can influence any data set, and add "these humble and human dilemmas of *real* life in *real* schools and communities that fly under the technocrat's radar" because "[o]n the path of technocracy, data are defined and operationalized narrowly, simplistically, and unthinkingly" (Hargreaves & Shirley, 2009, p. 33). And even worse, schools bought into it.

2.5.1.3 "The Path of Effervescence: Our Celebrations"

Once schools figured out how to work the accountability system to their advantage, they indulged in "[g]aming the system and savoring the public rewards for doing so" (Hargreaves & Shirley, 2009, p. 41). Principals are kissing pigs, camping out on rooftops, and handing out awards left and right. (In Texas, it is illegal to reward individual students for success on TAKS, but schoolwide celebrations where no one is

excluded/singled out are permissible.) We have rallies, field trip celebrations, "camps," and, most important of all, the public announcement of school rankings. These are heady things because they appear to convince schools and the larger communities that XYZ school is, in fact, *a good school*. We have bought into our own spin.

Hargreaves and Shirley (2009) call this "hyperactive professionalism," that—while heady and often exhilarating—is a distraction because it prevents schools from "developing and realizing inspiring purposes of their own" and instead focus on "energetically and enthusiastically delivering the government's narrowly defined targets and purposes" (p. 41). We embrace this illusion of proficiency and settle for this level of achievement rather than aiming higher.

The harm of this test-driven system has been noted elsewhere in this chapter, particularly in the work of Berliner and Nichols, and in Diane Ravitch's comments on the realities that NCLB produced. In closing the Third Way section, Hargreaves and Shirley (2009) succinctly state one of the greatest faults of the Third Way: "[I]t has not put the passion back into teaching nor the pleasure into learning" (p. 45).

2.5.1.4 Trapped in the Second Way

To put new ideas into practice would be to make a significant course correction in current educational policy at both the federal and state levels. Texas has a double dose of the philosophies behind NCLB, because before George W. Bush became president, he was governor of Texas and enacted many of what would go on to be key provisions of NCLB. Educators felt dishonored when Margaret Spellings, a political

science major with great political and bureaucratic experience but no educational experience and especially no classroom experience became Secretary of Education. If educators hoped for educational change with President Obama, to date they have been disappointed as the Obama administration is continuing with the high-stakes accountability testing programs of NCLB and even advancing them into determining teacher quality based on primarily (or even exclusively) on test results. Educational reforms are still inexorably linked with test scores from these state assessments. Obama's choice for Secretary of Education was Arne Duncan, whose claimed success as CEO of Chicago Public Schools is now under scrutiny, appears much the same as former Secretary of Education Rod Paige who entered office with accolades for being an architect of the "Texas Miracle" where urban students made rapid test gains only to see the "Miracle" debunked when researchers found the numbers to be manipulated by, among other things, misrepresenting the number of those who dropped out to avoid taking the test (*60 Minutes,* 2004). Duncan's claim of improvement is also under fire for closing neighborhood schools, which lead to an unintended but deadly increase in school violence (Rethinking Schools, 2011).

Again, education is an inherently complex construct and simple solutions will not address the problems faced. Ron Wolk (2011), founder and former editor of *Education Week,* writes that,

> In education we almost always cite research to buttress our opinions. I do in this book [*Wasting Minds*, 2011], even though I know that one can find some

> research to support almost any opinion. . . . I once believed that education research would lead us to the promised land of successful schools and high student achievement. I no longer believe that. (p. 13)

Wolk (2011) is brave enough to put into print what most educators often feel: there is evidence to support all conclusions. Much research uses test scores as indicators, but Wolk bluntly asserts "[i]f test scores are not a fair, accurate, or reliable measure of student learning, then the research and its conclusions are flawed and unreliable" (p. 130). And because there is "so much room for subjective interpretation of data in educational research, the findings often cancel each other out" (p. 13).

If, as Wolk (2011) claims, data is not to be trusted, then what is? Since test scores are the measurement of success ("the coin of the realm"), he includes them but also argues to make decisions based on "personal observations, logic, and common sense" (p. 130). This same combination should be permissible for all who need to make decisions and/or evaluations. It may be the only way out of the trap of Second Way thought.

Wolk (2011) asserts that *A Nation at Risk* "completely missed the real reasons for the poor performance of America's students and schools" by "not even mention[ing] poverty, race, urban schools, new immigrants, . . . the impact of popular culture" or the antiquated structures of schools (pp.17-18). Wolk considers this a wasted opportunity. The report garnered ground-breaking (for education) attention and motivation, but led

reformers into unproductive paths. They focused on "performance," and not "design." (Wolk is a proponent of Dennis Littky's Big Idea model.)

However, reformers and teachers can take some comfort in that education is cyclical. What has been before will be again. What is now will pass. Newkirk (2009) wrote of the early 20th century where high accountability in the form of testing was the sole measure of success. It passed and the reforms of John Dewey took place, and then the reforms occurred in various content areas like reading and writing workshop approaches and inquiry in science. Our current obsession with what Hargreaves and Shirley (2009) labeled "the unholy trinity: markets, testing and accountability" will pass as well (p. 25). This too, shall pass, but not without harm to both students and teachers. And not without harm to education and even our democracy. But maybe it will be relegated to the past permanently this time.

Moving forward with this study, my question is how do these national concerns relate to what has happened in Texas in the TAKS era and what will happen with the transition to STAAR? Are these issues as much of a concern here? Or have we managed to avoid the worst of it? My research looks to determine some quantitative data to see how prevalent some specific practices are.

2.6 The Fourth Way: Realistic Future or Fantasy?

The journey through 20th century educational reform can be summed up with the words of the *Grateful Dead* song: "What a long, strange trip it's been." We can also add exhausting, frustrating, expensive, and a host of other descriptors. It has been a

journey filled with solid progress and illusory gain, creative expression, and scientific attempts to chart that expression, multiple experiences of déjà vu, and overwhelming feelings of frustrations that things just are not right. A critical look at where we have been is essential before moving forward, and this chapter has thus far focused on the literature that details the issues that has led us to this point, but it is a meaningless exercise if we do not look to the future.

2.6.1 What Is Gold, What is Lead

Hargreaves and Shirley (2009) propose that the Fourth way should be a product of the three preceding periods, but that it should embrace the best of those periods and reject the things that soured the period. The best and the worst of the past six decades of educational change are summed up in the following table.

Each era had both positive and negative impacts on education. Hargreaves and Shirley (2009) suggest that educators and policymakers take the obvious approach and learn from what has come before. Their reform model calls for schools and communities to stand up and work beside government, with government, and in some cases, against government, to create the best possible schools (Hargreaves & Shirley, 2009).

Table 2.2 What to Retain and What to Abandon

	Retain	Abandon
First Way	Inspiration, innovation and autonomy	Inconsistency and professional license
Interregnum	Common standards with local interpretation	Weak development of teachers, leaders, and communities

Table 2.2--*Continued*

Second Way	Urgency, consistency and all-inclusive equity	Cut-throat competition and excessive standardization
Third Way	Balance and inclusiveness, public involvement, financial reinvestment, better evidence, and professional networks	Persistent autocracy, imposed targets, obsession with data, effervescent interactions.

Note. From *The Fourth Way: The Inspiring Future for Educational Change: The Inspiring Future for Educational Change,* by Hargreaves and Shirley, (2009, p. 48). Copyright 2009 by Corwin. Reprinted with permission. *Table 1 Hargreaves & Shirley, 2009.*

This is where education stands now, on the cusp of the Fourth Way. What will the course be? Another quick slide back into Second Way thinking? Or, truly and excitingly, a giant step forward? The answers are yet to be determined, and they—alas—are outside the scope of this study.

This review of literature has been lengthy, but the path of educational reform over the past half century has been complex, perplexing, and at times, illogical and counterproductive. It was my goal to explain how education arrived at the situation we find ourselves in regarding high-stakes testing, and especially the opportunities for the misuse of both tests and the data they produce. Chapter 3 explains the methodology of my research conducted to determine the extent of two such faulty practices: misuse of released tests (and the data produced from the tests) as benchmark testing and using any TAKS score alone as the basis for important decisions regarding students or instruction.

CHAPTER 3

METHODOLOGY

In chapter 2, I presented a detailed history of how American education has arrived at this juncture. High stakes testing is the law of the land because of NCLB and the desires of the current crop of school reformers who call for increasing accountability for schools, which translates as "more testing" regardless of what Berliner (2007) calls "collateral damage." It is in this environment that I conducted my research focusing on the use of released TAKS tests and TAKS scores in instructional settings. In chapter 3 I explain the methodology of my study, beginning with the development of my survey instrument.

3.1 Rationale

In chapter 1, I presented my three research questions. The survey was expected to produce descriptive data about how many districts in the sample were using released TAKS exams as benchmarks and how many used TAKS data alone (either from an actual TAKS administration or from the use of a released test used as a benchmark or combination of the two) to make important decisions such as placing students or evaluating instruction. Relying on inferential statistics to provide results, I developed hypotheses to determine if size of the school district could have an impact on the

prevalence of those practices. Sherman (2008) conducted preliminary research on this topic examining benchmarking in Texas school districts and presented her findings in an unpublished dissertation. Sherman's findings indicated that a majority of Texas school superintendents thought positively of the concept of benchmarking, and that 78% of school districts were using benchmarking. The composition of those benchmarks included released TAKS tests alone, or as a part of a combination of other resources (i.e., test preparation materials, teacher created materials, etc.). Sherman's research did not isolate the degree to which these released tests were used as the diagnostic data, so this study was built on her initial research to establish a quantitative source of data on that practice. I decided to use her model as my entry point for research. My research questions were a more precise and a critical extension of her questions to superintendents. Precise because my questions isolated the use of released TAKS instead of having them listed as a potential source of benchmark material and critical because as was pointed out in chapter 2, such a use goes against the test design and results in suspicious data, which is then used for instructional decisions.

Sherman (2008) surveyed school superintendents, but the actual composition of these benchmark tests were more familiar to the English/Language Arts supervisors or curriculum specialists within districts. In most cases, they were the ones responsible for creating these benchmarks or had direct contact with those who create them.

To find an answer to these questions, I conducted my research in two stages. The first stage was to develop an appropriate survey instrument that asked specific

questions about the composition of benchmark tests to produce usable data. The second was to actually administer the instrument to my sample group. The core of this sample was a professional organization of district-level ELA personnel. To increase the sample size, I contacted English/ Language Arts supervisors in geographic areas adjacent to the professional organization's locale to recruit participants.

3.2 Development and Validation of Survey Instrument

For ease and convenience of participants—and for its data generating capacities—an online survey was administered via SurveyMonkey™ to collect responses to the questions. After reviewing multiple survey instruments in the course of my literature review, I set up questions using frequency indicators (*always, frequently, occasionally, never*) as answer options. Again, after review, I developed 12 questions to cover the two research questions. (A copy of the preliminary survey is included in Appendix A.)

3.2.1. Determining Content Validity

After obtaining approval from my committee and the IRB, I determined content validity of the instrument utilizing procedures explained by Creswell (2005) and submitted my preliminary questions to a panel of experts for review. Since my questions dealt with district practices regarding use of released TAKS as benchmark/ diagnostic assessments, experts were those district-level personnel who know of the practice and its impact. In this case, district-level supervisors, specialists, and department chairpersons were recruited. I selected a group of approximately 15

educators from North Texas school districts (the same region represented by the core of the survey group, The professional organization) and asked them to review the questions and provide feedback, which was used to clarify the questions to achieve a more precise instrument. This was also the approach Sherman (2008) employed in her research involving district superintendents and benchmark testing, but interestingly, she did not conduct reliability testing of her survey instrument.

3.2.2 Determining Reliability

After amending the instrument in accordance with the suggestions of the content validity group, I selected a second group of approximately 15 different individuals from the same population (supervisors, specialists, and department chairpersons) to take the survey twice. Using the test-retest model explained by Cresswell (2005), I had the second group take the survey at two different intervals to determine reliability of the document. Thirteen respondents completed the test-retest model. After the second administration of the survey, I converted the responses (*always, frequently, sometimes, never*) into ordinal numbers (4, 3, 2, 1) and analyzed the data using SPSS.

Creswell (2005) recommended using a Spearman rho to establish reliability but other expert sources recommended a Pearson *r*. I ran both and found the results to be in agreement, with one exception that I discuss later. With a correlation of positive/negative 1 as the goal for reliability, the reliability for each question was established as

- Question 2: Pearson .836 and Spearman .788. The question was retained.

- Question 4: Pearson .394 and Spearman .369. The question was rejected.
- Question 5: Pearson .820 and Spearman .828. The question was retained.
- Question 6: Pearson .881and Spearman .820. The question was retained.
- Question 7: Pearson .751 and Spearman .734. The question was retained.
- Question 8: Pearson .910 and Spearman .930. The question was retained.
- Question 9: Pearson .814 and Spearman .826. The question was retained.
- Question 10: Pearson .034 and Spearman .051. The question was rejected.

One question, number 3, showed a larger difference between correlations. The Pearson correlation was .760 and the Spearman correlation was .517. After some consideration, the question was retained.

3.3 Participants and Calendar

As indicated, my representative core sample was the professional organization, The professional organization. This organization is comprised of reading and English supervisors representing schools in the Dallas-Fort Worth metropolitan area and surrounding counties. According to 2011-2012 student data from TEA, the student population of the schools in the sample pool was 625, 797. As supervisors, these members oversee ELA instruction in their respective districts and are knowledgeable about the content area as well as district practices. After reviewing the membership roster, I decided to increase the sample size by contacting English/Language Arts supervisors in districts that lay just beyond the professional organization boundaries. These educators were identified after reviewing the websites of those school districts in

Texas Educational Regional Service Centers' districts 10 and 11 and were contacted via email.

The survey was presented to the professional organization in March of 2012. Members who were not present at that meeting were contacted via email and asked to participate in the survey. All surveys were completed by March 31, 2012.

CHAPTER 4

PRESENTATION AND ANALYSIS OF DATA

4.1 Overview

The major purpose of this study was to survey the target group on their school districts' use of TAKS data (including released tests), and to determine if district type made a difference in how those districts utilized that information. The study also sought to supply quantitative data on the practice of using a released test as a benchmark measure. The three research questions were as follows:

1. What percentage of districts in the Dallas-Fort Worth area use TAKS data alone or as the major factor in making important decisions? To fully examine this question, I also examined if there was a difference in the answers from differing school compositions.
2. What percentage of districts in the Dallas-Fort Worth area use previously released TAKS exams as the instrument for benchmark testing? To fully examine this question, I also examined if there was a difference in the answers from differing school compositions.
3. What concerns do school districts have as Texas transitions from TAKS to STAAR?

SurveyMonkey™ supplied the vehicle for surveying the target group and compiling the data. I used descriptive statistics to analyze the data relating to the first and second research questions. I used inferential statistics, notably a Fisher's Exact test

in SPSS (version 19.0) computations to investigate the overarching area of questions 1 and 2 by testing a hypotheses for each item to determine if there was any statistical significance between the response and district type. I compiled and coded responses from Survey Item 9 to determine areas of concern about the transition from TAKS to STAAR.

The survey was sent to the 75 members in the target group. Of that number, 39, or 52%, completed the survey. This exceeds the amount generally agreed to be minimal for survey return rate, so I moved forward into analyzing the data. Table 4.1 shows how the respondents classified their school districts.

Table 4.1 Respondent Classifications

District Composition	Response
Urban	4.9%
Suburban	68.3%
Mid-Sized Town	17.1%
Rural	9.8%

The responses from four categories (*always, sometimes, seldom, never*) were collapsed to two categories (*always/sometimes* and *seldom/never*) and coded. Table 4.2 on the following page shows the results for the Survey Items 2-7. These results will be examined again shortly in more detail for Research Question 3.

There were also two open-ended response items to allow respondents to expand on any ideas or concerns. Response item 8 asked how TAKS had affected professional practice and had a response rate of 78.0%. Response item 9 asked what concerns the respondents had as Texas transitions from TAKS to STAAR, and the response rate for this item was 82.5%.

Table 4.2 Response Totals for Multiple Choice Item Responses 2-7

Survey Item	Percentage Responding Always/Sometimes	Percentage Responding Seldom/Never
2. Utilized released TAKS tests as benchmarks, diagnostic tests, and/or other instruments to prepare for actual TAKS administrations.	68.3%	31.7%
3. Administered released TAKS in "rehearsals" for TAKS that mimic the rules and procedures of an actual TAKS administration to the extent that the school schedule is significantly altered.	51.2%	48.8%
4. Produced dissagregated data (item analyses, population group performance, specific SE targets, etc.) to determine student strengths/weakness from benchmark administrations of released TAKS or TAKS formatted materials.	80.5%	19.5%
5. Utilized data from a released-test benchmark as the sole or major factor to determine such student interventions as placement in tutoring groups, remedial classes, special programs (e.g., Read 180), or reducing elective classes.	37.5%	52.5%
6. Based district-level instructional decisions solely or most heavily on the previous year's/years' TAKS scores, more so than grades, teacher/parent input, or	36.5%	63.4%

Table 4.2--*Continued*

other testing data.		
7. Based campus-level instructional decisions solely or most heavily on the previous year's/years' TAKS scores, more so than grades, teacher/parent input, or other testing data.	40%	60%

4.2 Results of Research Question 1

TAKS data includes the scores from both actual administrations and benchmark administrations, so Survey Items 4 through 7 provide information relevant to this question. (What percentage of districts in the Dallas-Fort Worth area use TAKS data alone or as the major factor in making important decisions?) Sherman (2008) found that 78% of school superintendents in her survey supported the practice of benchmarking, and the results of this survey were similar with 68.3% of respondents indicating their districts use released TAKS tests as benchmarks (Survey Item 2). These benchmarks produce data, which are harvested by 80.5% of representative districts (Survey Item 4). However, the responses to Survey Items 5, 6, and 7 showed that few school districts are making important decisions based solely on that data. Survey Item 5 specifically asked if the resultant data from a released TAKS benchmark is being used to make important decisions regarding students. Only slightly more than a third of respondents, 37.5%, indicated a positive response (*always* or *frequently*), with 62.5% indicating a negative response (*seldom* or *never*).

Survey Items 6 and 7 asked specifically about TAKS results being used to make important decisions at district and local levels. Again, the percentage of respondents indicating a positive response (*always* or *frequently*) was less than half, with 36.5% at

district level and 40% at the campus level. The majority of respondents rejected the idea that TAKS data alone was driving decision making, with 63.4% replying *seldom* or *never* at the district level and 60% replying likewise on campus level decisions.

The answers suggest that school districts use released TAKS as part of their benchmarking process and fully harvest data from their administrations. However, most do not use that data alone or as the major factor in making instructional decisions.

4.3 Results of Research Question 2

Survey Items 2, 3, and 4 are relevant to this question (What percentage of districts in the Dallas-Fort Worth area use previously released TAKS exams as instruments for benchmark testing?). Roughly two-thirds of the respondents (68.3%) indicated they used released exams as part of a benchmark/diagnostic assessment plan. About a quarter (24.4%) of the respondents indicated they occasionally used released tests, with only 7.3% indicating that they never utilized released exams. This result indicates that the use of released tests as part of a benchmarking/test preparation mode—although inappropriate to test design—is widespread.

Survey Item 3 addressed the use of these documents in a test rehearsal approach whereby the school day is significantly altered to resemble an actual testing situation. Over half of the respondents (51.2%) indicated that their school districts do this *always* or *frequently*. Of the remaining responses, 36.6% indicate that they have done this occasionally and only 12.3% indicate that they have *never engaged* in this practice.

Survey Item 4 addressed the results of the administration of these benchmark tests. A significant number (80.5%) reported they produced disaggregated data from those administrations. When combined with the number who reported doing so *occasionally*, that number rose to 98.6%. Only 1% reported never disaggregating the data. This indicates that districts are gathering from the misuse of a released TAKS test.

Survey Item 5 addressed how those results were used. Not quite half (47.5%) of the respondents said they always or frequently used this data as the sole or major criteria in determining student interventions. The slightly larger group, 52.5%, said they seldom or never used these results as the sole or major criteria for determining student interventions. The numbers in that larger group were close, with 25% saying they occasionally used the data alone/as a major factor and 27.5% saying they never used this data alone. (Given that the overwhelming number of respondents said they collected such data in Survey Item 4, it is noteworthy that over a quarter of respondents indicated they do not utilize the results of this administration in isolation.)

4.4 Results of Research Question 3

To determine the concerns teachers had regarding these issues, I asked two specific open-ended questions. Open-ended items were added to the survey to give teachers the opportunity to expand on the issues of the survey, but also to be considered as potential areas for further, more specific research and investigation. I coded the responses to Item 8 (what affect has TAKS had on professional practice) as either *positive, negative,* or *neutral.*

Table 4.3 Descriptors of TAKS Effect on Professional Practice

Response Category	Percentage Responding
Positive	32.25%
Negative	35.48%
Neutral	32.25%

The results were somewhat evenly divided. The two longest responses, however, were negative (373 words and 115 words). The average response was 35 words. The implications of these results will be discussed in chapter 5.

Response Item 9 asked what concerns the respondents had for the transition from TAKS to STAAR. The average response on this item was 36 words, with the longest (about teaching to the test) at 233 words and the second longest (also about teaching to the test) coming in at 130 words. A number of concerns were discussed and the responses were coded based on the topic of the concern. The most frequent concerns were rigor (27.0%), teaching to the test (21.0%), and a lack of specific information on the composition/scoring of the test from TEA (19.0%). The implications of these results will be discussed in chapter 5.

4.5 Hypotheses Developed To Test Relationship Between District Composition and Practices

While descriptive statistics were adequate to address the previous research questions, inferential statistics were needed for this analysis to examine if a school district's composition—urban, suburban, mid-sized town, or rural—had any

relationship to the district's use of released TAKS tests. These hypotheses were developed as follows:

- Hypothesis 1: There is no relationship between district composition and the use of released TAKS tests as benchmarks, diagnostic tests, and/or other instruments to prepare for actual TAKS administrations.
- Hypothesis 2: There is no relationship between district composition and the use of released TAKS as "rehearsals" for TAKS that mimic the rules and procedures of an actual TAKS administration to the extent that the school schedule is significantly altered.
- Hypothesis 3: There is no relationship between district composition and the use of dissagregated data (item analyses, population group performance, specific SE targets, etc.) to determine student strengths/weaknesses from benchmark administrations of released TAKS or TAKS formatted materials.
- Hypothesis 4: There is no relationship between district composition and the use of data from a released-test benchmark as the sole or major factor to determine such student interventions as placement in tutoring groups, remedial classes, special programs (e.g., Read 180), or reducing elective classes.
- Hypothesis 5: There is no relationship between district composition and basing district-level instructional decisions solely or most heavily on the previous year's/years' TAKS scores, more so than grades, teacher/parent input, or other testing data.

- Hypothesis 6: There is no relationship between district composition and basing campus-level instructional decisions solely or most heavily on the previous year's/years' TAKS scores, more so than grades, teacher/parent input, or other testing data.

I began by coding the cells of the responses to Survey Item 1 (district composition). Respondents who indicated they were either urban or suburban were assigned a 0, and respondents who indicated they were either mid-sized town or rural were assigned a 1. Likewise, I collapsed the cells for the responses to Survey Items 2 through 7, so that a positive response (*always* or *frequently*) was coded as 0, and a negative response (*occasionally* or *never*) was coded as 1. I then created a 2 x 2 Crosstab table.

I merged the data from SurveyMonkey™ into SPS. More specifically, I used Fisher's exact test because of the small sample size and developed a contingency table for each item analysis. Table 4.4 shows the number of respondents to each question.

Table 4.4 Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	*N*	Percent	*N*	Percent
Survey Item 2	39	100.0%	0	.0%	39	100.0%
Survey Item 3	39	100.0%	0	.0%	39	100.0%
Survey Item 4	39	100.0%	0	.0%	39	100.0%

Table 4.4--*Continued*

Survey Item 5	38	97.4%	1	2.6%	39	100.0%
Survey Item 6	39	100.0%	0	.0%	39	100.0%
Survey Item 7	38	97.4%	1	2.6%	39	100.0%

Hypothesis 1: There is no relationship between district composition and the use of released TAKS tests as benchmarks, diagnostic tests, and/or other instruments to prepare for actual TAKS administrations. Since the sample size was small and the expected frequencies were less than 5, a Fisher's exact test was used to test significance. No relationship was found between district composition and response, p = 1.0. The null hypothesis was not rejected. A phi coefficient was calculated to determine the magnitude of the relationship. According to Rea and Parker (1992) there is a negligible association between the two variables, phi = .010.

Table 4.5 Data Analysis of Survey Item 2 (Hypothesis 1)

		Item 2 Response		Total
		Always/ Frequently	Occasionally/ Never	
Urban/Suburban	Count	9	20	29
	Expected	8.9	20.1	29.0
Midsized Town/Rural	Count	3	7	10
	Expected	3.1	6.9	10.0
	Total	12	27	39

Hypothesis 2: There is no relationship between district composition and the use of released TAKS as "rehearsals" for TAKS that mimic the rules and procedures of an actual TAKS administration to the extent that the school schedule is significantly altered. Since the sample size was small and the expected frequencies were less than 5, a Fisher's exact test was also run to test significance. Since the sample size was small and the expected frequencies were less than 5, a Fisher's exact test was used to test significance. No relationship was found between district composition and response, p= .716. The null hypothesis was not rejected. A phi coefficient was calculated to determine the magnitude of the relationship. According to Rea and Parker (1992), there is a weak association between the two variables, phi = .102.

Table 4.6 Data Analysis of Survey Item 3 (Hypothesis 2)

		Item 3 Response		Total
		Always/ Frequently	Occasionally/ Never	
Urban/Suburban	Count	15	14	29
	Expected	14.1	14.9	29.0
Midsized Town/Rural	Count	4	6	10
	Expected	4.9	5.1	10.0
	Total	19	20	39

Hypothesis 3: There is no relationship between district composition and the use of dissagregated data (item analyses, population group performance, specific SE targets, etc.) to determine student strengths/weaknesses from benchmark administrations of released TAKS or TAKS formatted materials. Since the sample size was small and the expected frequencies were less than 5, a Fisher's exact test was also run to test significance. Since the sample size was small and the expected frequencies were less than 5, a Fisher's exact test was used to test significance. No relationship was found between district composition and response, $p = 1.0$. The null hypothesis was not rejected. A phi coefficient was calculated to determine the magnitude of the relationship. According to Rea and Parker (1992) there is a negligible association between the two variables, phi = .031.

Table 4.7 Data Analysis of Survey Item 4 (Hypothesis 3)

		Item 4 Response		Total
		Always/ Frequently	Occasionally/ Never	
Urban/Suburban	Count	5	24	29
	Expected	5.2	23.8	29.0
Midsized Town/Rural	Count	2	8	10
	Expected	1.8	8.2	10.0
	Total	7	32	39

Hypothesis 4: There is no relationship between district composition and the use of data from a released-test benchmark as the sole or major factor to determine such student interventions as placement in tutoring groups, remedial classes, special programs (e.g., Read 180), or reducing elective classes. Since the sample size was small and the expected frequencies were less than 5, a Fisher's exact test was used to test significance. No relationship was found between district composition and response , p = .269. The null hypothesis was not rejected. A phi coefficient was calculated to determine the magnitude of the relationship. According to Rea and Parker there is a moderate association between the two variables, phi = .239 .

Table 4.8 Data Analysis of Survey Item 5 (Hypothesis 4)

		Item 5 Response		Total
		Always/ Frequently	Occasionally/ Never	
Urban/Suburban	Count	16	12	28
	Expected	14.0	14.0	28.0
Midsized Town/Rural	Count	3	7	10
	Expected	5.0	5.0	10.0
	Total	19	19	39

Hypothesis 5: There is no relationship between district composition and basing district-level instructional decisions solely or most heavily on the previous year's/years' TAKS scores, more so than grades, teacher/parent input, or other testing data. Since the sample size was small and the expected frequencies were less than 5, a Fisher's exact test was also run to test significance. Since the sample size was small and the expected frequencies were less than 5, a Fisher's exact test was used to test significance. No relationship was found between district composition and response, $p = .141$. The null hypothesis was not rejected. A phi coefficient was calculated to determine the magnitude of the relationship. According to Rea and Parker (1992) there is a moderate association between the two variables, phi = .260.

Table 4.9 Data Analysis of Survey Item 6 (Hypothesis 5)

		Item 5 Response		Total
		Always/ Frequently	Occasionally/ Never	
Urban/Suburban	Count	20	9	29
	Expected	17.8	11.2	29.0
Midsized Town/Rural	Count	4	6	10
	Expected	6.2	3.8	10.0
	Total	24	15	39

Hypothesis 6: There is no relationship between district composition and basing campus-level instructional decisions solely or most heavily on the previous year's/years' TAKS scores, more so than grades, teacher/parent input, or other testing data. Since the sample size was small and the expected frequencies were less than 5, a Fisher's exact test was also run to test significance. Since the sample size was small and the expected frequencies were less than 5, a Fisher's exact test was used to test significance. No relationship was found between district composition and response, $p = .267$. The null hypothesis was not rejected. A phi coefficient was calculated to determine the magnitude of the relationship. According to Rea and Parker there is a moderate association between the two variables, phi = .217.

Table 4.10 Data Analysis of Survey Item 7

		Item 7 Response		Total
		Always/ Frequently	Occasionally/ Never	
Urban/Suburban	Count	18	10	28
	Expected	16.2	11.8	28.0
Midsized Town/Rural	Count	4	6	10
	Expected	5.8	4.2	10.0
	Total	22	16	38

CHAPTER 5

SUMMARY, FINDINGS, DISCUSSION, IMPLICATIONS FOR PRACTICE, AND RECOMMENDATIONS FOR FURTHER RESEARCH

In chapter 5, I summarize the scope and results of the study in six sections. In the first section, I present a summary of the study, which includes a look at the motivation for the study, the research questions, and the methods of the study. In the second section, I briefly encapsulate the findings of the study, which were presented in chapter 4. In the third section, I present a discussion of the study. In the fifth and sixth sections I examine the implications for educational practice and areas for further study, respectively.

5.1 Summary and Application

NCLB anchored American Education into the Second Way of Change by embracing standardization and accountability in the form of high-stakes testing, which in the State of Texas has meant TAKS. In this study I set out to find a quantitative measure of how widespread was the practice of utilizing released TAKS for diagnostic testing/benchmarking. It was, and is, a complaint often issued by teachers, and I sought to gain a measure of just how prevalent this practice was, as well as collecting information about the use of data harvested from such use.

In this study, I selected a sample group of English/language arts supervisors/ coordinators from the North Central Texas area. The core of this sample was a local affiliate of a statewide professional organization. The professionals in the sample group represented urban, suburban, small town and rural districts encompassing 625, 797 students. Fifty-two percent of participants responded to a survey of multiple choice and open-ended questions indicating the frequency of use of particular practices regarding released TAKS data (including released tests) via SurveyMonkey™ after establishing instrument validity and reliability as described in detail in chapter 3.

I gathered and tabulated results. Those results showed 68.3% of the respondents said their districts utilized released TAKS tests as benchmarks/diagnostic tests in preparation for the actual TAKS administration. Remember, that the TAKS technical manual as well as the concepts of formative and summative assessments do not support that usage. The data garnered from these administrations cannot be legitimately used for diagnostic purposes. That 68.3% of districts in the sample use released TAKS as benchmarks is worrisome because it indicates a large number of districts are not seeing the distinctions between summative and diagnostic assessment and the role of a benchmark test.

Slightly over half (51.2%) reported their districts significantly altered their school day to administer the benchmark. This shows that the complaints of teachers that too much time is giving over to such testing rehearsals are indeed justified. When taken with benchmarks in other content areas and considering that some schools give

benchmarks several times a year, the investment in time is considerable. It is important to note that time testing is time away from learning and time away from the engaging activities that pull students deeper into content.

Even more (80.5%) admitted that they used the released tests to generate disaggregated data about student performance. As indicated earlier, this data is highly suspect for diagnostic purposes. An excess of data also feeds into what Seife called "randumbness" or imagining patterns in data where none actually exist. (2011) The danger is that we diagnosis on shaky data and spend precious time on interventions that are little more than a shot in the dark. Too many students have been diverted into remedial programs to address perceived weaknesses based upon benchmark performance while ignoring the real issues that could not be measured on such a test.

Roughly a third of the respondents (ranging from 36.5% to 40%) indicated that TAKS data was the sole or major factor in making important decisions regarding student interventions and campus/district decisions. This number was a pleasant surprise for me, and I do believe that had this research taken place a few years earlier the number would have been higher. This shows that schools are becoming more aware of the importance of multiple measures in making decisions. However, that is still a large number of students who are being funneled into remedial classes, tutoring groups, and other interventions based on imprecise information. Nichols and Berliner's notion of collateral damage from high-stakes testing applies here.

A Fisher's Exact Test was run on the data from the multiple choice responses to determine if there was any significance between the answers of those professionals in larger districts (urban/suburban) or smaller districts (small town/rural). Few systems are monolithic, and school systems are known for variance. Larger districts have more resources to create curriculum and assessments, and I wanted to see if this made a difference in the practices regarding those released tests. Across the board, it did not.

The two open-ended optional questions allowed the survey pool to add their voices and comment on the effect of TAKS on their practice and their concerns as Texas phases out TAKS and moves to STAAR. Those who answered the question regarding the effect of TAKS on their practice were almost evenly split among positive, neutral, and negative feelings. A large number, however, had negative feelings as we make the move to STAAR.

These results are interesting. The question asking the affect of TAKS on their practice was almost evenly divided among the three categories of positive, negative, and neutral. It is rather like a Rorschach Test where different viewers will see different things. We could note that only a third of the respondents felt TAKS had positively impacted their practice, or take the opposite position that only a third of respondents felt that TAKS had negatively impacted their practice. The group that responded neutrally is puzzling: did they answer neutrally because they just ignored all the "fuss 'n holler" of the TAKS era and follow their own internal teaching compass, or was it because they felt the TAKS era truly had no effect? That is an area for further research.

The goal of my research and specifically my research questions was to determine a quantitative number for these practices in the state of Texas. My study provided those, and showed the numbers justify the concern that teachers have expressed. The results of my study fill in the gap between an understanding of the issues of standardized testing (as explored in Chapter 2) and the actual use of such tests in Texas. This research adds to the conversation and suggests additional research to extend understanding.

5.2 Discussion

Testing is a massive construct. NCLB made high-stakes testing the major—and often only—indicator of academic progress. Reams of paper and incalculable bytes of digital information have been given to reporting how tests are formed, delivered, and results dissected. Policymakers debate test scores, news outlets announce the USA is underperforming other nations (debatable, certainly). The scope is so big that how this plays out in classrooms is often overlooked. In this study I wanted to reveal some quantitative numbers for one practice within that massive mechanism of testing: how do schools in Texas utilize released TAKS tests (that term in itself is redundant since the A in TAKS stands for assessment, yet TAKS alone does not specify that I am referring to the actual released test document).

From my research as part of my Review of Literature, I knew that using such a document as a diagnostic test or benchmark was not an acceptable use according to the test design explained in the technical manual for the test (TEA, 2007). Furthermore,

whenever I met with teachers locally or at statewide meetings, I kept hearing that schools were devoting large amounts of time to benchmarking and administering these released tests in large scale reenactments of an actual test administration, some lasting as long as an entire day for one test. I earnestly wanted a quantitative number to apply to this practice and therefore this study was enacted. I knew from Sherman's (2008) research that a large number of district superintendents supported the concept of benchmarking, and while the idea of benchmarking/giving a diagnostic test may be sound, the quality of the benchmark is linked to the quality of the data harvested. A released TAKS test is a summative document and does not meet the criteria for a diagnostic benchmark (Popham, 2009). As such, any data harvested from it and used to determine student interventions is questionable at best, and more than likely invalid.

This study showed that in the target group 68.5% were using released tests as their benchmark assessment in preparation for the actual TAKS administration. Even more, however, used the released tests in other formats to gather information about student performance. While the group was admittedly a sample, the sample population still worked with 625,797 students. Therefore, generalizations can be made for Texas as a whole, and cautious generalizations can be made from these results for other states who utilize a similar testing framework.

Since the time I began this study, the landscape of testing has been changing. The near unanimous acclaim for data-driven applications exists no longer as questions about the validity and use of standardized tests have surfaced. In the spring of 2012,

proponents of testing suffered setbacks with testing debacles in New York (a questionable passage and questions went viral over the Internet and earned the sobriquet "Pineapplegate"; Haimson, 2012) and Florida (the state's FCAT writing scores were so low they exited the realm of possibility and revealed a faulty test design; Farley, 2012) brought intense scrutiny. During the first five months of 2012, more than 500 Texas school boards signed a resolution calling for the scaling back of standardized testing (Texas Association of School Administrators, 2012). The voices of those quoted in this document who assert the tests do not measure what is claimed have gained more prominence.

The title of this dissertation comes from an old story from one of my favorite books as a child. A man walking down a street one night notices another man on his hands and knees beneath a streetlight. The first man asks the second man what he is doing and that man replies he is looking for his keys. The first man joins in the search. After some time of fruitless search, the first man asks, "Are you sure you dropped your keys here?" and the second man responds that no, he had dropped them further down the block. The first man then asks why on earth he is looking at this spot and the second man replies, "Because the light's better here."

Too often with testing—including the use of those released tests as benchmarks—we will admit the tests are not the best, but they are all we claim to have. The second man never found his keys, and we will never really understand student achievement because our viewing area—testing—is too narrow and not focused on

where learning really is. Rothstein wrote that, "measurement of student achievement is complex—too complex for the social science methods presently available" (as cited in Harris, Smith, & Harris, 2011, p. 34). The "hidden costs" of these types of assessments are "enormous" (Harris et al., 2011, p. 111). In their book *The Myths of Standardized Tests: Why They Don't Tell You What You Think They Do* (2011), Harris, et al. used the same story of the man searching for the keys to explain why we find the tests so attractive and why they, ultimately, do not deliver on their promises. Even worse, the way test scores are used causes harm. Harris et al. write, "We believe great harm is done by treating the results, which are gross and fuzzy indicators at best, as if their validity is absolute, their apparent precision is real, and their unexamined outcomes unchallengeable" (p. 63). Their research confirms that of Berliner and Nichols (2007) referenced in chapter 2 and, once again, points to the collateral damage such an emphasis on testing produces.

5.3 Findings

I examined three research questions and an overarching area for research questions 1 and 2 in this study.

- Research Question 1 asked what percentage of districts in the North Central Texas area use TAKS data alone or as the major factor in making important decisions? Using descriptive statistics, I used the survey data to determine that respondents indicated that the number using TAKS data alone to make important decisions was only 36.6%. On the one hand, that is encouraging

information because it means almost two-thirds of districts are using multiple sources of data, as is appropriate. However, it means that fully 1 in 3 districts in the sample is making important decisions based on highly suspect data. The harms discussed in Chapter 2 are playing out in at least this many districts in the sample group.

- Research Question 2 asked what percentage of districts in the North Central Texas area use previously released TAKS exams as the instrument for benchmark testing? Again, descriptive statistics showed that 68.5% were using the released tests as benchmarks—in violation of the test design. Again, a large population is using released tests in an inappropriate manner, so any action taken based on the results of those benchmarks is highly suspect. With the stakes so high on misuse of such data, this is a worrisome result.
- Research Question 3 asked what concerns do school districts have as Texas transitions from TAKS to STAAR? Several areas of concern emerged from an analysis of the open-ended questions, especially with regard to concern about the supposed rigor of the test and lack of specific information from TEA regarding the composition and scoring of STAAR. It is important to note that even though this study focused on TAKS, it is still relevant for the STAAR era. Many of the practices associated with TAKS were carry-overs from the TAAS era, and it is very likely the same will occur with this

transition also. Much will be the same: released tests, high stakes, benchmarking, etc., so this study will still inform the discussion.

- The overarching area for research questions 1 and 2 asked if there was a relationship between district composition and the use of TAKS data, including the results of a released TAKS benchmark. Through the use of inferential statistics described in chapter 4, results showed there was no relationship between the composition of the district and the use of these practices. A hypothesis was generated for each item on the survey that related to this question and then analyzed with SPSS, notably a Fisher's exact test. These results showed that the faulty uses of released TAKS and TAKS data are not specific to any specific school composition. Large or small, urban or rural, the practices are the same.

As indicated earlier in this document, the goal of this study was to establish a quantitative measure for practices regarding the use of TAKS data that concern teachers in Texas. The study confirmed that a large number of school districts are indeed using released TAKS tests as benchmark instruments, and many in large-scale rehearsals that mimic an actual administration. The validity of such usages is questionable.

Since TAKS (and its replacement, STAAR) are developed to be summative exams, their use as diagnostic exams is inappropriate. Diagnostic testing requires multiple items on a single concept to determine student learning, a summative test evaluates only a representative sample of student learning. Using a summative tool as a

diagnostic can generate faulty data. This study found that 68.5% of school districts were engaging in this practice.

However, the number of districts that utilize TAKS data alone (TAKS data in this study was defined as information pulled from both actual and benchmark administrations) was noticeably lower at 36.6%. While there is still concern that roughly one in three districts is using only one data source for important decisions regarding students. It is heartening to see that approximately two-thirds of districts are using multiple sources of data to evaluate student learning.

Composition of district—urban, suburban, small town or rural—made no difference in the use of these practices, indicating that the effects of TAKS are somewhat uniform across the state. Larger districts with multiple resources or curriculum staff and assessment departments utilize the released documents in the same manner as small schools without such resources.

While I was working on this dissertation, resistance to standardized testing grew at an impressive rate. Many of the works cited in this study, notably Ravitch's book, helped fuel a movement that has exploded into the national dialogue on educational reform with advocates such as FairTest.org, the Opt Out movement, Parents Across America, and others demanding that America take a closer look at the testing behemoth. In the spring of 2012 two events validated their concerns.

The first event was dubbed "Pineapplegate" (Hartocolis, 2012). Eighth graders in New York took their reading test and then came home and hit the Internet with the

tale of the talking pineapple who challenged a hare to a race. The passage was idiotic and the possible answers were more so. They found other students in other states who had experienced the pineapple story in years past, and tracked down the author of the piece. The media picked up on the story and interviewed the author who explained the piece had ever been nonsensical and was originally used to demonstrate the unreliability of the speaker. Pearson quickly announced that the passage and its questions would not be used in determining student scores, but the damage was done (Haimson, 2012; Hartocolis, 2012). People began to ask, just what is in those tests? One parent citing the high-stakes nature of the tests (both student and teacher success would be determined by the results) filed a Freedom of Information Act to have the test released for scrutiny (Albin, 2012).

The second event happened in Florida, where the writing scores took such a precipitous dip that the results were out of the realm of possibility. As Rita Solnet, one of the founders of Parents Across America expressed, "[l]ast week we learned that children could not write any better than martians arriving on the planet yet later in the week the news trumpeted that Florida's reading scores improved across the state." (Solnet, 2012) The same students who had such dismal writing scores were reading successes. Finally people started asking what was happening. Who was guaranteeing the accountability of the tests? Farley's book *Making the Grade* had exposed the scoring practices, and he was on hand again to say, "I told you so" (Farley, 2012).

When I began this study, the role of standardized test data in education reform was unchallenged, and while there are still those who cling to the concept that the high-stakes testing data provides the measure for student learning and teacher effectiveness, a more balanced approach is beginning to develop. Bill Gates has backed off the exclusive use of test scores to evaluate teachers (Gates, 2012). Robert Scott, head of TEA has called testing a "perversion" (Jones, 2012). Parents across the nation are beginning to question the worth (and harm) of these tests.

The results of this study come at an interesting time for Texas teachers: the spring of 2012 saw the first administration of STAAR, and this study, I believe, raises some interesting implications for practice. How can we apply what we have learned in the TAKS era to bring forward the best of our understanding and leave behind the practices that were harmful or not effective?

5.4 Implications for Practice

1. Addressing "abysmal assessment literacy." As discussed in chapter 2, Popham (2009) coined this phrased to describe how educators utilize data without fully understanding it, such as using a released TAKS test for a benchmark exam or comparing TAKS scores to previous years. This study showed that a large number of districts are engaging in activities with released TAKS tests and TAKS data that do not indicate assessment literacy, particularly related to an understanding of the appropriate uses of each. Texas, and by extension America, needs to make the education of school

administrators in understanding data a priority. We have done an admirable job in teaching schools to use data, but a terrible job in teaching them to understand the data and its limitations. Books studies for administrators (and perhaps teachers) over Koretz's *Measuring Up* (2008) and Seife's *Proofiness* (2010) would be good places to begin.

2. Understanding why any future released STAAR exam needs to be handled the same way as a released TAKS, that is, not as a diagnostic. TEA has stated we will not have a released STAAR exam until 2014 (after the third administration). The basic ideas about assessment—formative, summative, diagnostic, etc.—need to be better understood by both classroom teachers and those responsible for accountability. With better assessment literacy, those involved could better understand the instruments used to assess and determine more appropriate uses. For example, if a school wanted to run a rehearsal for STAAR by letting students see the length of the test, the types of questions, and so forth, a dress rehearsal would be an acceptable use. Such a practice could also test school resources and diagnose problems with the actual administration. However, any data harvested from such a test can only show what students would have done had this been the test form, had this been the test day. It cannot reveal academic weaknesses (because it did not assess all academic areas sufficiently) and it cannot predict how the students will perform (because the actual test form could have a very

different composition; the practice was only a proxy). If a school chose to dedicate time to a dress rehearsal, those school administrators need to carefully weigh the loss of instructional time with the pseudo test. Is it really worth it? This study has shown that schools are using released tests in violation of test design, so these concerns are very real.

3. Developing an adequate system of diagnostic assessments. A main criticism of high-stakes testing is that it is one test on one day. A different way to address the measurement of student learning would be to create a series of interim assessments, or true benchmarks, of student progress throughout the year. The best of these, as explained by Dufour's (2004) model, calls for local development at the campus level by teachers in a Professional Learning Community (PLC) model. This gives ownership and flexibility to those most involved in the process. However, I think it would be helpful for TEA to release sample interim assessments to guide a process whereby school districts could create their own assessments. Again, this study has shown how summative documents are being used for diagnostic purposes in many districts, indicating, perhaps, a desperation on the part of those schools. For TEA to take a more proactive role in distinguishing between summative and diagnostic assessments would be very helpful.
4. Continued pressure on the testing industry for transparency. The testing industry has ballooned into a massive, secretive, and expensive enterprise

thanks to NCLB. As mentioned in chapter 2, Texas has a near half billion dollar contract with Pearson for the STAAR assessments. More information from the testing companies regarding test creation, composition, and scoring must be made available to warrant trust in Pearson's ability to adequately assess student learning. Such information needs to be presented in an accessible manner as well, and not hidden behind a wall of statistical jargon. Popham (2009) specifically addressed that, leveling condemnation at those who manipulated mathematics to keep critics at bay. Laws prohibit teachers from discussing the test with their students or even reading the test document. These laws should be revisited. Accountability is a two-way street. This study showed that a large number of districts are very accepting of released tests: they use them as benchmarks and use the data generated to make decisions (either alone or with other metrics). Schools—and parents—should become critical consumers of these products, and continue to press the testing companies for more transparency and accountability.

5. Development of multiple measures. At the time of this writing in the spring of 2012, many Texas school districts are applying to be part of a consortium of schools that will be given special freedoms from state requirements as a part of Senate Bill 1557. One of the major areas of that project involves developing a system of multiple measures of assessment for accountability instead of a single high-stakes score. While the consortium schools are only

a small number of the total Texas schools, they are paving the way for the acceptance of new approaches in Texas education. This concept is not new—it goes back to Third Way approaches and our brief experiences with them in the days before NCLB. This study showed that fully a third of respondents were not utilizing multiple assessments even though the TAKS Technical Manual called for the use of multiple assessments in making any sort of decision about students or instruction.

5.5 For Further Research

1. What are the practices developing for STAAR prep? As we move from the TAKS to the STAAR era, research areas will become apparent. Based on the results of this study, there is a need to examine benefit/harm of particular prep activities. What is being deleted from the school day/year to provide time for practice tests? How do students and teachers feel about these prep activities? Do campuses that avoid a heavy test prep mentality score as well as those who do intensive test prep? Also, how possible is it that students could score well on the test, yet still be deficient in skills? Likewise, the reverse: Can competent students score poorly on the tests? When Texas transitioned from TAAS to TAKS, many practices and materials were carried forward. Will that happen again? Very probably. This study showed the extent of such practices as benchmarking with released tests and using TAKS data to determine

interventions. As we make the move to STAAR, there needs to be careful consideration of what we take with us and what we leave behind.

2. What are the degrees of assessment literacy among administrators and teachers? Do they really understand what the data does/does not indicate? This is a very interesting area. Few—I think—administrators and teachers would admit they do not understand the data they deal with on a daily basis. A carefully constructed study that utilized data scenarios and applications could be quite revealing. I want to be clear that I do not fault anyone for not understanding the complexities of data. Most teachers and administrators were presented the data-driven agenda in a very simple form and led to believe that the number is all that matters. Popham's (2009) ideas about educator proficiency with data need a quantitative base.
3. What is the extent of test composition in determining curriculum? In the new world of STAAR, how close are teachers staying to the standards? Are they adhering to all the major areas, or privileging the ones that are tested? For example, the English I STAAR/EOC calls for two compositions. Do teachers focus on those two tested genres and ignore the genres not tested that year? If so, what does that create for teachers in later years who are dealing with students who will be tested on those genres? Are teachers still saying they cannot teach as they would like to—by which I mean doing what they feel is the best for their students—because of perceptions about the test? This

study was always viewed as foundational, that is, it would provide quantitative data that would raise more questions. This conversation will continue and subsequent researchers will add to it.

4. A closer look at test composition and the workings of Pearson and other testing companies. As referenced earlier in this chapter, it has not been a good spring for Pearson. Specific research studies addressing the composition of those exams and the validity and reliability of their scoring need to be conducted independent of the company itself. While their internal research is adequate to guide internal decisions, people independent of the company must be able to access their information and conduct research in order for there to be a true trust in the testing process. This study revealed that these tests—both through released tests as benchmarks and test data—have enormous impact upon students and teachers. In that case, shouldn't we have the best possible instruments?

5.6 Conclusion

In my opinion, there is a need in the State of Texas and across America to better understand the high-stakes testing documents and scores. I began this study with the idea that there were more test scores floating around the educational landscape than there was understanding of those scores. The New Cuyama sign was emblematic of the problem as it showed valid numbers used in an invalid manner. The study ends with a

confirmation of that idea and evidence to establish the specific mischance, misdirection and misapplication of that data. The New Cuyama sign is still an appropriate symbol.

APPENDIX A

COPY OF THE SURVEY INSTRUMENT

Please identify the size of your school district

5A 4A 3A 2A 1A

Please identify the type of district

urban suburban mid-sized town rural

Please select the extent to which schools in your district have utilized the following practices:

1. Utilizing entire released TAKS tests as benchmarks, diagnostic tests, or other documents to prepare students for the actual TAKS test.

always frequently occasionally seldom never

2. Administering released TAKS in so-called "rehearsals" for TAKS or "testing mode" wherein the enter grade/school mimics the rules of an actual TAKS scenario and disrupts the school day for the administration.

always frequently occasionally seldom never

3. Producing disaggregated data (item analysis, population group data, data regarding specific TEKS student objectives within subjects, etc.) from the administration of released tests.

always frequently occasionally seldom never

4. Producing limited basic data (pass/fail, indentify by subject--i.e. reading or writing) to determine student weakness.

always frequently occasionally seldom never

5. Utilizing the data from a released-test benchmark as the sole or major factor to determine student interventions (tutoring groups, enrollment in remedial classes or special programs such as Read 180, eliminating electives, etc).

always frequently occasionally seldom never

6.Utilizing sections of released tests in TAKS preparation efforts.

always frequently occasionally seldom never

7. Utilizing the previous year's/years' TAKS scores as the sole or most important factor (more so than grades, teacher/parent input, other testing data, etc.) in making instructional decisions at a district level.

always frequently occasionally seldom never

7. Utilizing the previous year's TAKS scores as the sole or most important factor (more so than grades, teacher/parent input, other testing data, etc.) to determine student interventions (tutoring groups, enrollment in remedial classes or special programs such as Read 180, eliminating elective courses, etc.).

always frequently occasionally seldom never

APPENDIX B

PERMISSIONS

Copyright Clearance Center

Confirmation Number: 10995007
Order Date: 05/02/2012

Customer Information

Customer: Ramona Lowe
Account Number: 3000529238
Organization: Ramona Lowe
Email: lower@lisd.net
Phone: +1 (948)4698216
Payment Method: Credit Card ending in 5237

Order Details

The fourth way : the inspiring future for educational change

Billing Status: **Charged to Credit Card**

Order detail ID:	62442023	**Permission Status:**	**Granted**
ISBN:	978-1-4129-7637-4	**Permission type:**	Republish or display content
Publication year:	2012	**Type of use:**	Republish in a dissertation
Publication Type:	Book	**Requested use:**	Dissertation
Publisher:	Corwin Press	**Republication title:**	"But the Light's Better Here," He Said: Misdirection, Mischance and Misapplication in Data-Driven Applications
Rightsholder:	SAGE PUBLICATIONS INC BOOKS	**Republishing organization:**	University of Texas at Arlington
Author/Editor:	Andy Hargreaves and Dennis Shirley	**Organization status:**	Non-profit 501(c)(3)
Your reference:	Ramona's dissertation, chapter 2	**Republication date:**	08/01/2012
		Circulation/ Distribution:	10
		Type of content:	Chart
		Description of requested content:	Figure 2.1
		Page range(s):	44
		Translating to:	No Translation
		Requested content's publication date:	05/02/2012
		Payment Method:	CC ending in 5237

Rightsholder terms apply (see terms and conditions) **$ 3.50**

Total order items: 1 **Order Total: $3.50**

THE UNIVERSITY
OF TEXAS
AT ARLINGTON

Office of Research Administration
Box 19188
202 E. Border St., Suite 214
Arlington, Texas
76019-0188

T 817.272.3723
F 817.272.1111

http://www.uta.edu/research
Expertise at UT Arlington
http://www.uta.edu/expertise

February 03, 2012

Ramona Lowe
Dr. Jeanne Gerlach
College of Education and Health Professions
Box 19277

Protocol Title: *Validation of Survey Instrument for TAKS Released Test Usage*

RE: Exempt Approval Letter

IRB No.: **2011-0617e**

The UT Arlington Institutional Review Board (UTA IRB) Chair (or designee) has reviewed the above-referenced study and found that it qualified as exempt from coverage under the federal guidelines for the protection of human subjects as referenced at Title 45 Part 46.101(b)(1)(2). You are therefore authorized to begin the research as of January 20, 2012.

Please be advised that as the principal investigator, you are required to report local adverse (unanticipated) events to this office within 24 hours. In addition, pursuant to Title 45 CFR 46.103(b)(4)(iii), investigators are required to, "promptly report to the IRB ***any*** proposed changes in the research activity, and to ensure that such changes in approved research, during the period for which IRB approval has already been given, are **not initiated without IRB review and approval** except when necessary to eliminate apparent immediate hazards to the subject."

All investigators and key personnel identified in the protocol must have documented Human Subject Protection (HSP) Training or *CITI Training* on file with this office. The UT Arlington Office of Research Administration Regulatory Services appreciates your continuing commitment to the protection of human research subjects. Should you have questions or require further assistance, please contact Robin Dickey at robind@uta.edu or you may contact the Office of Regulatory Services at 817-272-3723.

Sincerely,

Patricia G. Turpin, PhD, RN, NEA-BC
Clinical Associate Professor
UT Arlington IRB Chair

REFERENCES

Abydos Learning. (n.d.). *Organization information.* Retrieved from. http://www.Abydoslearning.org

Adler, B. (2010, November 23). Collaborative learning. *Newsweek*. Retrieved from http://www.newsweek.com/2010/11/23/ collaborative-learning.html

Albin, J. (2012, May 11). A request to make the Pearson tests public. *New York Times Schoolbook.* Retrieved from http://www.nytimes.com/schoolbook/2012/05/11/a-request-to-make-the-pearson-tests-public/

Allington, R. (Ed.). (2002). *Big brother and the national reading curriculum: How ideology trumped evidence*. Portsmouth, NH: Heinemann.

Allington, R. (2008). *What really matter in response to intervention: Researched based design.* Needham Heights, MA: Allyn & Bacon.

American Educational Research Association. (2000). *High-stakes testing in pre K-12 education*. Retrieved from http://www.aera.net

Ansary, T. (2007, March). Education at risk: Fallout from a flawed report. [Electronic version]. *Edutopia*. Retrieved from http://www.edutopia.org/landmark-education-report-nation-risk

Ariel Investments (2011). *Our history*. Retrieved from http://www.arielinvestments.com/ourhistory/

Atwell, N. (1987). *In the middle*. Portsmouth, NH: Boynton-Cook.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36, 258-267.

Azzam, W. (2008). Learning about—and from—data. *Educational Leadership*, 66, 91-92.

Bambrick-Santoyo, P. (2010). *Driven by data: A practical guide to improve instruction.* San Francisco, CA: Jossey-Bass.

Bracey, G. (2006). *Reading educational research: How to avoid getting statistically snookered*. Portsmouth, NH: Heinemann.

Brown, J., Gutstein, E., Lipman, P. (2009, Spring) Rethinking Schools. Retrieved from http://www.rethinkingschools.org/restrict.asp?path=archive/23_03/arne233.shtml

Burke, G. (2012). *Henry Louis Mencken*. Retrieved from http//:www.io.com/gibbonsb/mencken/

Bushaw, W., & Lopez, S. (2011, August 17). Betting on teachers: The 43rd annual Phi Delta Kappa/Gallup poll of the public's attitude toward schools. *Phi Delta Kappan*. Retrieved from http://www.pdkintl.org/poll/docs/pdkpoll43_2011.pdf

Callahan, R. (1962). *Education and the cult of efficiency*. Chicago, IL: University of Chicago Press.

Charlotte-Mecklenburg Public Schools. (2011). Talent effectiveness project. Retrieved fromhttp://www.cms.k12.nc.us/cmsdepartments/accountability/payforperformance/Pages/default.aspx

Chavez, S. (2009, March 25). Grand Prairie High School principal singles out black students to improve TAKS scores. *Dallas Morning News*. Retrieved from Denton Record-Chronicle http://www.dentonrc.com/sharedcontent/ dws/dn/latestnews/ stories/032509dnmetgpprincipal.37afb68.html

Chronicle of Higher Education. (2008, June 28). *Writing is the key to college success. Do your students write well right now?* Retrieved from http://chronicle.com/ article/Writing-is-the-Key-to-College/46940/

Creswell, J. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Saddle River, NJ: Pearson.

Crook, S. (2007). *Item analysis*. Retrieved from http://www.drshirleycrook.com.

Cronin, J., Dahlin, M., Adkins, D., & Kingsbury, G. (2007). *The proficiency illusion*. Washington, D.C.: The Fordham Foundation.

Dickens, C. (1859). *Tale of two cities*. Retrieved from http://etext.virginia.edu/toc/ modeng/public/DicTale.html

Dillion, S. (2010, December 7). *Top test scores from Shanghai stun educators.* Retrieved from http:nytimes.com/2010/12/07/education /07education.html

Dufour, R., Eaker, R., Karhanek, G., & Dufour, R. (2004). *Whatever it takes: How professional learning communities respond when students don't learn.* Bloomington, IN: Solution Tree.

Egan, K. (2005). Testing what for what? *Educational Leadership*, 63, 27-30.

Farley, T. (2012, May 20). For profit standardized testing can't be trusted. *Tampa Bay Times*. Retrieved from http://www.tampabay.com/opinion/columns/for-profit-standardized-testing-industry-cant-be-trusted/1230845

Farley, T. (2009). *Making the grade: My misadventures in the standardized testing industry.* Sausalito, CA: PoliPoint Press.

Felch, J., Song, J., & Smith, D. (2010, August 14). *Who's teaching L.A.'s kids? Los Angeles Times*. Retrieved from http://www.latimes.com/news /local/la-me-teachers-value-20100815,0,2695044.story

Feller, B. (2006, September 23). *Audit finds ethical lapses in U.S. reading program.* Retrieved from http://libproxy.uta.edu:2236/iwsearch/we/InfoWeb?p_product=AWNB&p_theme=aggreated5&p_action=doc&p_docid=114543A281DDBCC8&p_docnum=2&p_queryname=2.

Florio, M. (2010, December 28). Debate continues regarding Vikings-Eagles postponement. *NBC Sports: ProFootball Talk.* Retrieved from www.profootballtalk.nbcsports.com/2010/12/28/debate-continues-regarding-eagles-vikings-postponement/related/

Fox, J. (2008, October 23). The comeback Keynes. *Time.* Retrieved from http://www.time.com/time/magazine/article/ 0,9171,1853302,00.html

Fusarelli, L. (2004, January). The potential impact of the No Child Left Behind Act on equity and diversity in American education. *Educational Policy*, 18. Retrieved from http://epx.sagepub.com/content/18/1/71.short

Gilliam, F. (1999). *The welfare queen experiment: How viewers react to images of African-American mothers on welfare*. Cambridge, MA: Harvard University Nieman Reports. Retrieved from www.nieman.harvard.edu/ reportsitem.aspx?id=102223

Glod, M. (2009, June 23). The schoolhouse flunks. *Washington Post*. Retrieved from www.washingtonpost.com/wp-dyn/content/article/ 2009/06/22/ AR2009062202971.html

Graves, D. (2002). *Testing is not teaching: What should count in education*. Portsmouth, NH: Heinemann.

Guggenheim, D. (Writer/Director). (2010) *Waiting for Superman* [documentary]. Hollywood, CA: Paramount.

Hacker, H. (2010, March 21). Students playing catch up as they hit college. *Dallas Morning News.* Retrieved from http://www.dallasnews.com/news/ education/headlines/30100320-students-playing-catch up-as-they-4288.ece

Hargreaves, A., & Shirley, D. (2009). *The fourth way: The inspiring future for educational change.* Thousand Oaks, CA: Corwin Press.

Harris, P., Smith, B., & Harris, J. (2011). *The myths of standardized testing: Why they don't tell you what you think they do.* Lanham, MD: Rowan & Littlefield.

Harvey, J. (2004). The matrix reloaded. *Educational Leadership*, 61, 18-22.

Hess, F. (2008). The new stupid. *Educational Leadership* 66, 12-17.

Hess, F. (2010). *The same thing over and over: How school reformers get stuck in yesterday's ideas.* Boston, MA: Harvard University Press.

Hill, J. (photographer) 2007. *Sign's wrong* [photograph]. Retrieved from Flickr at http://www.flickr.com/photos/ theprimaryjosh/2150953626

Hortocolis, A. (2012, April 20) When pineapple races hare, students lose, critics of standardized testing say. *New York Times*.

Hout, M., & Elliot, S. ed. (2011) *Incentives and test based accountability in education*. Washington, D.C.: National Academies Press.

Jacobs, A. (2010, October 6). Rampant fraud threat to China's brisk ascent. *New York Times*. Retrieved from http://www.nytimes.com/2010/10/07 /world/asia/ 07fraud.html?_r=4 &pagewanted=1

Jerald, C. (2003). Beyond the rock and the hard place. *Educational Leadership*, 61, 12-17.

Johnson, J. (2007). What does the public say about accountability? *Educational Leadership*, 61, 36-41.

Kilgo, M. (n.d.). *TAKS analysis.* Retrieved from www.margaretkilgo.com

Klein, A. (2011, March 14). Obama calls for NCLB fix, warns against education cuts. *Education Week*. http://www.edweek.org/ew/articles/2011/03/ 14/26obama.h30.html

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice*, 22, 18-26.

Koretz, D. (2009). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press.

Kwan, Y. (2010). Life satisfaction and self-assessed health among adolescents in Hong Kong. *Journal of Happiness Studies*, 3, 383-393.

Lazear, E. (2006). Speeding, terrorism, and teaching to the test. *The Quarterly Journal of Economics,* 121, 1029-1061. Retrieved from http://qje.oxfordjournals.org /content/121/3/1029.short

Leung, R. (Reporter) (2004, August 25). The Texas miracle. *60 Minutes*. Transcript retrieved from http://www.cbsnews.com/stories/2004/01/06/6

Longman, J., & Macur, J. (2008, July 27). Records say Chinese gymnasts may be underage. *New York Times*. Retrieved from www.nytimes.com/2008/07/27/ sports/olympics/ 27gymnasts.html

Lyons, D. (2010, December 20). Bill Gates and Randi Weingarten [Interview]. *Newsweek.* Retrieved from http://www.newsweek.com/2010/12/20/gates-and-weingarten-fixing-our-nation-s-schools.html

Manzo, K. (2008, October 15). Latest reading first study reports limited benefits. *Education Week. http://www.edweek.org/ew/articles/2008/10 /15/08reading.h28.htm*

Marzano, R. (2003). *What works in schools: Translating research into action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Mathis, W. (2010). *The common core standards initiative: An effective reform tool?* Boulder University of Colorado at Boulder: Education and the Public Interest Center.

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

Mellon, E. (2010, June 7). Qualms arise over TAKS standards. *Houston Chronicle*. Retrieved from http://www.chron.com/disp/story.mpl /metropolitan/ 7041445.html

Merrow, J. (2011, May 26). The international education divide. *Huffington Post.* Retrieved from http://www.huffingtonpost.com/john-merrow/international-education-divide_b_867393.html

MGA Entertainment. [Website] (n.d.). Ownership information. Retrieved 1/1/11 from http://www.mgae.com/international.asp

Morsy, L., Kieffer, M., & Snow, C. (2010). Measure for measure: A critical consumer's guide to reading comprehension assessments for adolescents. *A time to act.* New York, NY: Carnegie Corporation of New York's Council on Advancing Adolescent Literature.

MSNBC. (2010, December 7). Wake up call: U.S. students trail global leaders. Retrieved from http://www.msnbc.msn.com/id/40544897/ns/us_news-life/t/wake-up-call-us-students-trail-global-leaders/#.TkgO8HMkcfk

MSNBC. (2011, April 7). *NYC schools chancellor Cathie Black Quits.* Retrieved from http://www.msnbc.msn.com/id/42474860/ns/us_news/t/nyc-schools-chancellor-cathie-black-quits/

Murray, D. (1972). Teaching writing as a process not a process, not a product. In V. Villanueva (Ed.), *The leaflet cross-talk in comp theory* (2003). Urbana, IL: National Council of Teachers of English.

National Board for Professional Teaching Standards. (2011). Retrieved from http://www.nbpts.org

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Retrieved from http://www2.ed.gov/pubs/NatAtRisk/index.html

National Writing Project. (2011). *History.* Retrieved from http://www.nwp.org.

Newkirk, T. (2009). *Holding on to good ideas in a time of bad ones: Six literacy principles worth fighting fo*r. Urbana, IL: National Council of Teachers of English.

Nichols, S., & Berliner, D. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard University Press.

Olson, K. (2009). *Wounded by school: Recapturing the joy in learning and standing up to old school culture.* New York, NY: Teachers College Press.

Pink, D. (2009). *Drive: The surprising truth about what motivates us*. New York, NY: Riverhead Publishing.

Popham, W. (2008). Anchoring down the data. *Educational Leadership*, 66, 85-86.

Popham, W. (2009). *Unlearned lessons: Six stumbling blocks to our schools' success*. Cambridge, MA: Harvard University Press.

Popham, W. (2003). *Test better, teach better: The instructional role of assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

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

Rea, L.M. & Parker, R.A. (1992) *Designing and conducting survey research*. San Francisco: Jossey Bass.

Reeves, D. (2008) Looking deeper into the data. *Educational Leadership*, 66, 89-90.

Reynolds, J. (2008, September 10). *Chinese youth 'face suicide risk.* BBC. Retrieved from http://news.bbc.co.uk/2/hi/asia-pacific/7608575.stm

Robbins, A. (2007). *The overachievers: The secret lives of driven kids*. New York, NY: Hyperion.

Robelen, E. (2011, January 26). *Missouri district competes against nations. Education Week.* Retrieved from http://www.edweek.org/ew/ articles/2011/01/26/ 18pisa.h30.html

Rothstein, R. (2009). What's wrong with accountability by the numbers? *American Educator*, *33*(1), 20-33.

Sacks, P. (1999) *Standardized minds: The high price of America's testing culture and what we can do to change it.* New York, NY: Da Capo.

Samuels, C. (2011, August 10). Cheating scandals intensify focus on test pressures. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2011/08/04 /37cheating_ep.h30html?r=1588658185

Santayana, G. (n.d.) [Quote]. Retrieved from http://www.saidwhat.co.uk/quotes/ favourite/ george_santayana

Schmoker, M. (2008). Measuring what matters. *Educational Leadership*, 66, 70-74.

Schmoker, M. (2006). *Results now: How we can achieve unprecedented improvements in teaching and learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

Seife, C. (2010). *Proofiness: The dark arts of mathematical deception*. New York, NY: Viking Press.

Sherman, A. (2008). *Preparing for the Texas Assessment of Knowledge and Skills (TAKS): A study of local benchmark testing in public schools.* Doctoral Dissertation. Retrieved from Pro Quest. UMI 3318640

Smith, J. (2011, March 2011). Texas ISD [website]. Retrieved from www.texasisd.com Cost of STAAR

Solnet, R. (2012, May 22). *Statistically speaking, Florida, I don't believe them.* Retrieved from http://www.dailykos.com/story/2012/05/22/1093715

Sports Illustrated. (2011, January 18). New information in the case against Lance Armstrong. Retrieved from http://sportsillustrated.cnn.com/2011/ more/01/18 /lance.armstrong/ index.html.SI. Com is listed as author

Stack, M. (2011, January 13). Chinese students' high scores in international tests come at a cost. *Los Angeles Times*. Retrieved from http://www.latimes.com/news/ nationworld /world/la-fg-china-education- 20110113,0,6192691.story

Stahl, L. (2010, September 29). Laura Bush expected to unveil major public school reform program, an initiative of SMU's George W. Bush Institute. *Dallas Morning News.* Retrieved from http://www.dallasnews.com/news/ education/headlines/20100928-Laura-Bush-expected-to-unveil-major-3481.ece

Stedman, L. (1993). The Sandia Report and U.S. Achievement. *The Journal of Educational Research*, 87, 133-146.

Texas Education Agency. (2007). Technical digest for the academic year 2007-2008. Retrieved from http://www.tea.state.tx.us/student.assessment/techdigest/yr0708/

U.S. Department of Education. (2008). *Reading First impact study.* Retrieved from ies.ed.gov/ncee/pdf/20084016.pdf. This was a published study from the DOE

U.S. Department of Education. (2011). *Arne Duncan biography.* Retrieved from http://www2.ed.gov/news/staff/bios/duncan.html

U.S. Department of Veterans Affairs. (2011). *GI Bill history*. Retrieved from http://www.gibill.va.gov/GI_Bill_Info/history.htm.

Vu, P. (2008, January 17). *Do state tests make the grade?* Retrieved from http://www.stateline.org/live/details/story?contentid=2772382

Wadhwa, V. (2011, January 12). U.S. schools are still ahead—way ahead. *Businessweek.* Retrieved from www.businessweek.com/technology/content /jan2011/ tc20110112_006501.htm

Weiss, J. (2012, February 28). Group of successful schools criticizes state testing system. *Dallas Morning News,* 1B, 7B.

William, D. (2010). What counts as evidence of educational achievement? The role of constructs in the pursuit of equity in assessment. *Review of Research in Education*, 34, 254-284

Wolk, R. (2011). *Wasting minds: Why our education system is failing and what we can do about it*. Alexandria, VA: Association for Supervision and Curriculum Development.

Wolk, R. (2009). Why we're still at risk: The legacy of five faulty assumptions. *Education Week*, 28, 30-36.

Zavis, A., & Barboza, T. (2010, September 28). Teacher's suicide shocks school. *Los Angeles Times.* Retrieved from http://articles.latimes.com/2010/sep/28/local/la-me-south-gate-teacher-20100928

Zhao, Y. (2009). *Catching up or leading the way: American education in the age of globalization.* Alexandria, VA: Association for Supervision and Curriculum Development.

BIOGRAPHICAL INFORMATION

Ramona Lowe was born in Wurzburg, Germany, and grew up as an "army brat," attending seven schools in her six years of elementary attendance before her father retired in Lawton, Oklahoma. She credits the amazing English teachers she had at Tomlinson Junior High and Eisenhower High School during the marvelous First Way Era for channeling her passion for reading and writing.

She attended Cameron University, earning a B.A. in English and later earned a master's degree in English Education from Southwestern Oklahoma State University. Later, she completed the hours needed to earn Reading Specialist certification. She was active in the Oklahoma Council of Teachers of English, serving as president in 1995-96. She was also on the board of directors for the Lawton Area Reading Council. Her proudest accomplishment was earning National Board Certification (as an Early Adolescent Generalist) in 1998.

After 19 years of teaching in Oklahoma, she moved to Texas to take a position at a large suburban high school. In 2005, she took a position in a neighboring district, first working on an interdisciplinary team based at an at-risk middle school and later in the district curriculum office.

Away from school she is a devoted aunt to Rachel and Mark, a breast cancer survivor, dog lover, and long-time Anglophile.